PASW® Data Preparation (Vorbereitung von Daten) 18

Weitere Informationen zu SPSS Inc.-Software-Produkten finden Sie auf unserer Website unter der Adresse *http://www.spss.com* oder wenden Sie sich an

SPSS Inc.
233 South Wacker Drive, 11th Floor
Chicago, IL 60606-6412, USA
Tel.: (312) 651-3000
Fax: (312) 651-3668

SPSS ist eine eingetragene Marke.

PASW ist eine eingetragene Marke von SPSS Inc..

Die SOFTWARE und die Dokumentation werden mit BESCHRÄNKTEN RECHTEN zur Verfügung gestellt. Verwendung, Vervielfältigung und Veröffentlichung durch die Regierung unterliegen den Beschränkungen in Unterabschnitt (c)(1)(ii) von The Rights in Technical Data and Computer Software unter 52.227-7013. Vertragspartner/Hersteller ist SPSS Inc., 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.
Patentnr. 7.023.453

Allgemeiner Hinweis: Andere in diesem Dokument verwendete Produktnamen werden nur zu Identifikationszwecken genannt und können Marken der entsprechenden Unternehmen sein.

Windows ist eine eingetragene Marke der Microsoft Corporation.

Apple, Mac und das Mac-Logo sind Marken von Apple Computer, Inc., die in den USA und in anderen Ländern eingetragen sind.

Dieses Produkt verwendet WinWrap Basic, Copyright 1993–2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Kein Teil dieser Publikation darf ohne vorherige Einwilligung des Herstellers in irgendeiner Form elektronisch, mechanisch, durch Fotokopieren, Aufzeichnen oder sonstige Mittel reproduziert, in einem Datenabfragesystem gespeichert oder übertragen werden.

Vorwort

PASW Statistics 18 ist ein umfassendes System zum Analysieren von Daten. Das optionale Zusatzmodul Data Preparation (Vorbereitung von Daten) bietet die zusätzlichen Analyseverfahren, die in diesem Handbuch beschrieben sind. Die Prozeduren im Zusatzmodul Data Preparation (Vorbereitung von Daten) müssen zusammen mit PASW Statistics 18 Core verwendet werden. Sie sind vollständig in dieses System integriert.

Installation

Zur Installation von Data Preparation (Vorbereitung von Daten) Erweiterungsmodul führen Sie den Lizenzautorisierungsassistenten mit dem Autorisierungscode aus, den Sie von SPSS Inc. erhalten haben. Weitere Informationen finden Sie in den Installationsanweisungen im Lieferumfang von Data Preparation (Vorbereitung von Daten) Erweiterungsmodul.

Kompatibilität

PASW Statistics kann auf vielen Computersystemen ausgeführt werden. Mindestanforderungen an das System und Empfehlungen finden Sie in den Unterlagen, die mit Ihrem System geliefert werden.

Seriennummern

Die Seriennummer des Programms dient gleichzeitig als Identifikationsnummer bei SPSS Inc.. Sie benötigen diese Seriennummer, wenn Sie sich an SPSS Inc. wenden, um Informationen zum Kundendienst, zu Zahlungen oder zu Aktualisierungen des Systems zu erhalten. Die Seriennummer wird mit dem Core-System ausgeliefert.

Kundendienst

Wenden Sie sich bei Fragen zur Lieferung oder Ihrem Kundenkonto an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Halten Sie bitte stets Ihre Seriennummer bereit.

Ausbildungsseminare

SPSS Inc. bietet öffentliche und unternehmensinterne Seminare an. Alle Seminare beinhalten auch praktische Übungen. Seminare finden in größeren Städten regelmäßig statt. Wenn Sie weitere Informationen zu diesen Seminaren wünschen, wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden.

Technischer Support

Kunden mit Wartungsvertrag können den Technischen Support in Anspruch nehmen. Kunden können sich an den Technischen Support wenden, wenn sie Hilfe bei der Arbeit mit PASW Statistics oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Informationen zum Technischen Support finden Sie auf der Website unter *http://www.spss.com* oder wenden Sie sich an Ihr regionales Büro, das Sie auf der Website unter *http://www.spss.com/worldwide* finden. Bei einem Anruf werden Sie nach Ihrem Namen, dem Namen Ihrer Organisation und Ihrer Seriennummer gefragt.

Weitere Veröffentlichungen

Das Handbuch *SPSS Statistics Statistical Procedures Companion* von Marija Norušis wurde von Prentice Hall veröffentlicht. Eine neue Fassung dieses Buchs mit Aktualisierungen für PASW Statistics 18 ist geplant. Das Handbuch *SPSS Statistics Advanced Statistical Procedures Companion*, bei dem auch PASW Statistics 18 berücksichtigt wird, erscheint demnächst. Das Handbuch *SPSS Statistics Guide to Data Analysis* für PASW Statistics 18 wird ebenfalls derzeit erstellt. Ankündigungen für Veröffentlichungen, die ausschließlich über Prentice Hall verfügbar sind, finden Sie auf der Website unter *http://www.spss.com/estore* (wählen Sie Ihr Land aus und klicken Sie auf Books).

Inhalt

Teil I: Benutzerhandbuch

1 Einführung in Data Preparation (Aufbereitung von Daten) 1

Verwendung der Prozeduren von “Data Preparation” (Vorbereitung) 1

2 Validierungsregeln 2

Vordefinierte Validierungsregeln laden 2
Validierungsregeln definieren 3
Definieren von Regeln für eine Variable 4
Definieren von Regeln für mehrere Variablen 6

3 Daten validieren 8

Daten validieren: Grundlegende Prüfungen 10
Daten validieren: Regeln für eine Variable 12
Daten validieren: Regeln für mehrere Variablen 13
Daten validieren: Ausgabe 14
Daten validieren: Speichern 15

4 Automatisierte Datenaufbereitung 17

So rufen Sie die automatische Datenaufbereitung ab 19
So rufen Sie die interaktive Datenaufbereitung ab 19
Registerkarte “Felder” 20
Registerkarte “Einstellungen” 20
Datum und Uhrzeit aufbereiten 21
Felder ausschließen 22
Messniveau anpassen 23
Datenqualität verbessern 24
Felder neu skalieren 25

Felder transformieren . . . 26
Auswählen und erstellen . . . 27
Namensfelder . . . 28
Transformationen anwenden und speichern . . . 29
Registerkarte “Analyse” . . . 31
Feldverarbeitungsübersicht . . . 32
Felder . . . 33
Aktionsübersicht . . . 35
Vorhersagekraft . . . 36
Feldertabelle . . . 37
Felddetails . . . 38
Aktionsdetails . . . 40
Transformiert Werte zurück . . . 43

5 Ungewöhnliche Fälle identifizieren 45

Ungewöhnliche Fälle identifizieren: Ausgabe . . . 47
Ungewöhnliche Fälle identifizieren: Speichern . . . 48
Ungewöhnliche Fälle identifizieren: Fehlende Werte . . . 49
Ungewöhnliche Fälle identifizieren: Optionen . . . 50
Zusätzliche Funktionen beim Befehl DETECTANOMALY . . . 51

6 Optimales Klassieren 52

Optimales Binning – Ausgabe . . . 54
Optimales Binning – Speichern . . . 55
Optimales Binning – Fehlende Werte . . . 56
Optimales Binning – Optionen . . . 57
Zusätzliche Funktionen beim Befehl OPTIMAL BINNING . . . 58

Teil II: Beispiele

7 Daten validieren 60

Validieren einer medizinischen Datenbank . . . 60
Durchführen von grundlegenden Prüfungen . . . 60
Kopieren und Verwenden von Regeln aus einer anderen Datei . . . 64

Definieren von eigenen Regeln . . . 73
Regeln für mehrere Variablen . . . 79
Fallbericht . . . 80
Zusammenfassung . . . 80
Verwandte Prozeduren . . . 81

8 Automatisierte Datenaufbereitung 82

Interaktive Verwendung der automatisierten Datenaufbereitung . . . 82
Auswahl aus Objekten . . . 82
Felder und Felddetails . . . 90
Automatische Verwendung der automatisierten Datenaufbereitung . . . 95
Vorbereitung der Daten . . . 95
Erstellen eines Modells mit unvorbereiteten Daten . . . 99
Erstellen eines Modells mit den vorbereiteten Daten . . . 102
Vergleichen der Vorhersagewerte . . . 104
Rücktransformieren der Vorhersagewerte . . . 105
Zusammenfassung . . . 107

9 Ungewöhnliche Fälle identifizieren 108

Algorithmus für "Ungewöhnliche Fälle identifizieren" . . . 108
Identifizieren ungewöhnlicher Fälle in einer medizinischen Datenbank . . . 109
Durchführen der Analyse . . . 109
Zusammenfassung der Fallverarbeitung . . . 113
Liste der Indizes anomaler Fälle . . . 114
Liste der Gruppen-IDs anomaler Fälle . . . 115
Liste der Gründe anomaler Fälle . . . 116
Normwerte der metrischen Variablen . . . 117
Normwerte der kategorialen Variablen . . . 118
Auswertung des Anomalie-Index . . . 120
Auswertung der Gründe . . . 120
Streudiagramm des Anomalie-Index über den Variableneinfluss . . . 121
Übersicht . . . 123
Verwandte Prozeduren . . . 124

10 Optimales Klassieren 125

Der Algorithmus für optimales Klassieren . . . 125

Verwenden der optimalen Klassierung zur Diskretisierung der Daten zu Kreditantragstellern . . . 125
Durchführen der Analyse . . . 126
Deskriptive Statistiken . . . 129
Modellentropie . . . 130
Klassierungs-Zusammenfassungen . . . 131
Klassierte Variablen . . . 135
Anwenden von Syntax-Klassierungsregeln . . . 135
Zusammenfassung . . . 137

Anhang

A Beispieldateien 139

Bibliografie 152

Index 154

Teil I: Benutzerhandbuch

Kapitel

1

Einführung in Data Preparation (Aufbereitung von Daten)

Der Informationsbedarf wächst proportional mit dem Anstieg der Leistungsfähigkeit von Computern. Das führt zu immer größeren Datensammlungen, zu mehr Fällen, mehr Variablen und mehr Fehlern bei der Dateneingabe. Diese Fehler behindern Vorhersagen auf der Grundlage von Prognosemodellen, dem wichtigsten Ziel des Daten-Warehousing. Deswegen müssen die Daten "sauber" gehalten werden. Die Menge der gespeicherten Daten ist jedoch bereits so weit über die Kapzitäten zur manuellen Prüfung der Daten hinausgewachsen, dass es entscheidend ist, automatisierte Prozesse für die Datenvalidierung zu implementieren.

Mit dem Erweiterungsmodul "Data Preparation" (Aufbereitung von Daten) können Sie ungewöhnliche und ungültige Fälle, Variablen und Datenwerte im aktuellen Datenblatt identifizieren und Daten zur Modellierung vorbereiten.

Verwendung der Prozeduren von "Data Preparation" (Vorbereitung)

Es hängt von Ihren Bedürfnissen ab, welche Prozeduren von "Data Preparation" (Vorbereitung) für Sie infrage kommen. Nachdem Sie die Daten geladen haben, könnte eine typische Vorgehensweise folgendermaßen aussehen:

- **Vorbereitung der Metadaten.** Überprüfen Sie die Variablen in der Arbeitsdatei, und bestimmen Sie die gültigen Werte, Labels und Messniveaus. Identifizieren Sie die Kombinationen von Variablenwerten, die zwar unmöglich, jedoch häufig falsch kodiert sind. Definieren Sie auf der Grundlage dieser Informationen Validierungsregeln. Dies kann zeitraubend sein, ist jedoch den Aufwand wert, wenn Sie regelmäßig Datendateien mit ähnlichen Attributen validieren müssen.
- **Datenvalidierung.** Führen Sie grundlegende Prüfungen und Prüfungen mit definierten Validierungsregeln durch, um ungültige Fälle, Variablen und Datenwerte zu identifizieren. Wenn sie ungültige Daten gefunden haben, untersuchen und beseitigen Sie die Ursache. Dies macht möglicherweise einen weiteren Durchlauf durch die Vorbereitung der Metadaten erforderlich.
- **Vorbereitung des Modells.** Verwenden Sie die automatisierte Datenvorbereitung, um Transformationen der ursprünglichen Felder zu erhalten, die die Modellerstellung verbessern. Identifizieren Sie potenzielle statistische Ausreißer, die in vielen Vorhersagemodellen Probleme verursachen können. Einige Ausreißer sind das Ergebnis von ungültigen Variablenwerte, die noch nicht identifiziert wurden. Dies macht möglicherweise einen weiteren Durchlauf durch die Vorbereitung der Metadaten erforderlich.

Sobald die Datendatei "sauber" ist, können Sie Modelle in anderen Erweiterungsmodulen erstellen.

Kapitel
2

Validierungsregeln

Eine Regel wird verwendet, um zu entscheiden, ob ein Fall gültig ist. Es gibt zwei Typen von Validierungsregeln:

- **Regeln für eine Variable.** Regeln für eine Variable bestehen aus einer festen Gruppe von Tests, die auf eine einzige Variable angwendet werden, z. B. Tests auf Werte außerhalb des Bereichs. Bei den Regeln für eine Variable können die gültigen Werte als Wertebereich oder als eine Liste zulässiger Werte ausgedrückt werden.
- **Regeln für mehrere Variablen.** Regeln für mehrere Variablen stellen benutzerdefinierte Regeln dar, die auf eine einzige Variable oder eine Kombination von Variablen angewendet werden können. Regeln für mehrere Variablen bestehen aus einem logischen Ausdruck, der ungültige Werte kennzeichnet.

Die Validierungsregeln werden im Datenlexikon Ihrer Datendatei gespeichert. Dies ermöglicht es, die Regeln einmal zu definieren und später wiederzuverwenden.

Vordefinierte Validierungsregeln laden

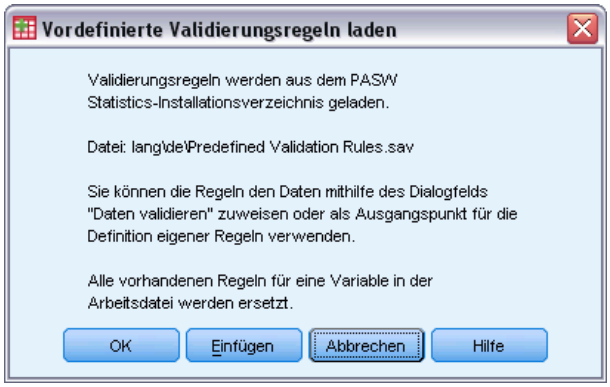
Sie können schnell auf eine Gruppe gebrauchsfertiger Validierungsregeln zugreifen, indem Sie vordefinierte Validierungsregeln aus einer externen Datendatei laden, die in der Installation enthalten ist.

So laden Sie vordefinierte Validierungsregeln:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Daten
  Validierung
    Vordefinierte Regeln laden...

Abbildung 2-1
Vordefinierte Validierungsregeln laden


Vordefinierte Validierungsregeln laden
Validierungsregeln werden aus dem PASW Statistics-Installationsverzeichnis geladen.
Datei: lang\de\Predefined Validation Rules.sav
Sie können die Regeln den Daten mithilfe des Dialogfelds "Daten validieren" zuweisen oder als Ausgangspunkt für die Definition eigener Regeln verwenden.
Alle vorhandenen Regeln für eine Variable in der Arbeitsdatei werden ersetzt.
OK
Einfügen
Abbrechen
Hilfe

Beachten Sie, dass hierbei alle vorhandenen Validierungsregeln für eine Variable in der Arbeitsdatei gelöscht werden.

Sie können auch den Assistenten zum Kopieren von Dateneigenschaften verwenden, um Regeln aus einer beliebigen Datendatei zu laden.

Validierungsregeln definieren

Im Dialogfeld “Validierungsregeln definieren” können Sie Validierungsregeln für eine oder mehrere Variablen erstellen und anzeigen.

So erstellen Sie Validierungsregeln und lassen diese anzeigen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:

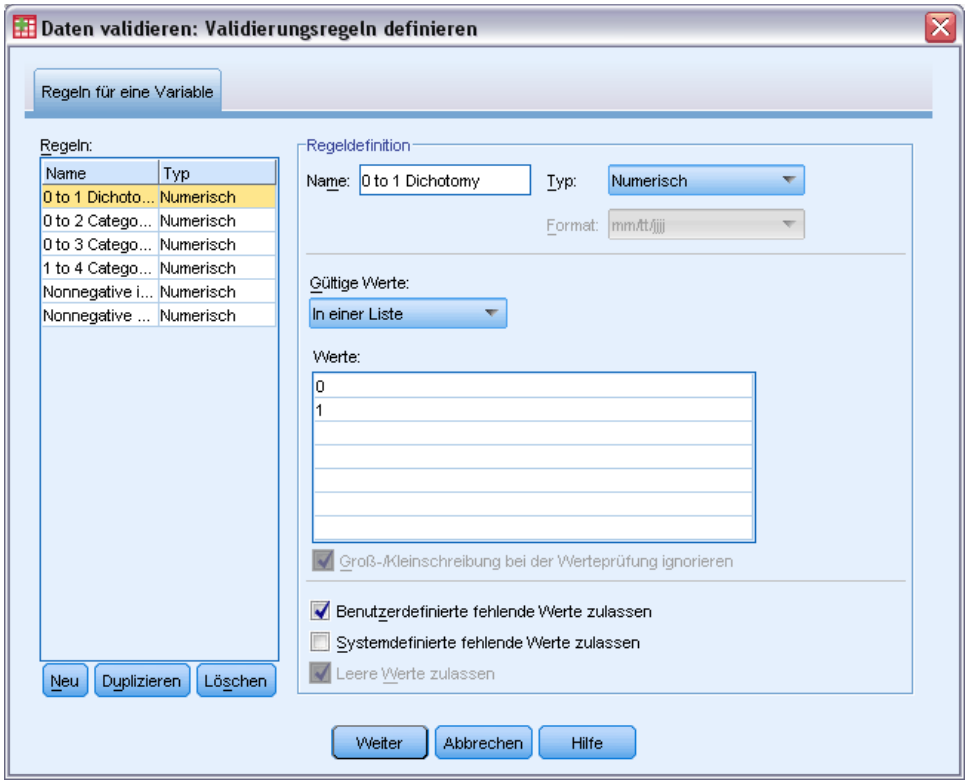
  Daten
  Validierung
  Regeln definieren...

  Das Dialogfeld wird mit Validierungsregeln für eine oder mehrere Variablen ausgefüllt, die aus dem Datenlexikon ausgelesen werden. Wenn keine Regeln vorliegen, wird automatisch eine neue Regel als Platzhalter erzeugt, die Sie nach Bedarf anpassen können.

- Wählen Sie einzelne Regeln auf den Registerkarten “Regeln für eine Variable” und “Regeln für mehrere Variablen” aus, um sich die Eigenschaften anzeigen zu lassen und diese zu ändern.

Definieren von Regeln für eine Variable

Abbildung 2-2
Dialogfeld "Validierungsregeln definieren", Registerkarte "Regeln für eine Variable"



Auf der Registerkarte "Regeln für eine Variable" können Sie Validierungsregeln für eine Variable erstellen, anzeigen lassen und ändern.

Regeln. Die Liste zeigt die Validierungsregeln für eine Variable nach Namen und Variablentyp, auf den die jeweilige Regel angewendet werden kann. Wenn Sie das Dialogfeld öffnen, werden die im Datenlexikon definierten Regeln angezeigt. Falls gegenwärtig keine Regel definiert ist, wird eine Platzhalter-Regel mit dem Namen "EinVarRegel 1" angezeigt. Unter der Liste "Regeln" werden folgende Schaltflächen angezeigt:

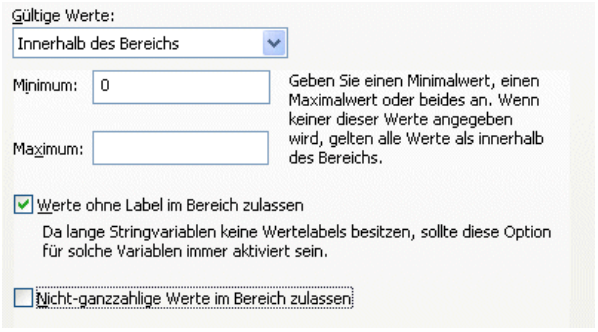
- **Neu.** Fügt einen neuen Eintrag am Ende der Liste "Regeln" hinzu. Die Regel wird ausgewählt und erhält den Namen "EinVarRegel *n*". Hierbei ist *n* eine Ganzzahl, sodass der Name der Regel unter den Regeln für eine oder mehrere Variablen eindeutig ist.
- **Duplizieren.** Fügt eine Kopie der ausgewählten Regel am Ende der Liste "Regeln" hinzu. Der Name der Regel wird so angepasst, dass er unter den Regeln für eine oder mehrere Variablen eindeutig ist. Wenn Sie beispielsweise "EinVarRegel 1" duplizieren, erhält die erste duplizierte Regel den Namen "Kopie von EinVarRegel 1", die zweite den Namen "Kopie (2) von EinVarRegel 1" usw.
- **Löschen.** Löscht die ausgewählte Regel.

Regeldefinition. Mit diesen Steuerelementen können Sie die Eigenschaften für eine ausgewählte Regel anzeigen lassen und festlegen.

- **Name.** Der Name der Regel muss unter den Regeln für eine oder mehrere Variablen eindeutig sein.
- **Typ.** Dies ist der Variablentyp, auf den die Regel angewendet werden kann. Wählen Sie Numerisch, String oder Datum aus.
- **Format.** Hiermit können Sie das Datumsformat für die Regeln auswählen, die auf Datumsvariablen angewendet werden können.
- **Gültige Werte.** Sie können die gültigen Werte als Bereich oder als Werteliste angeben.

Mit den Steuerelementen zum Festlegen eines Bereichs können Sie einen Bereich gültiger Werte angeben. Werte, die sich außerhalb dieses Bereichs befinden, werden als ungültig gekennzeichnet.

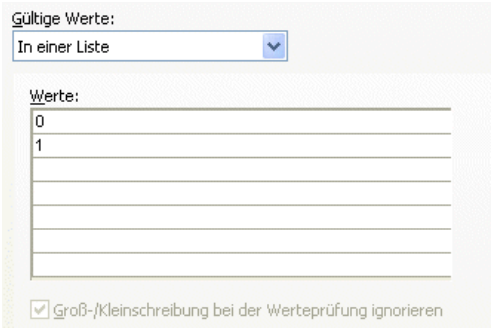
Abbildung 2-3
Regeln für eine Variable: Bereichsdefinition


Gültige Werte:
Innerhalb des Bereichs
Minimum:
0
Maximum:
Geben Sie einen Minimalwert, einen Maximalwert oder beides an. Wenn keiner dieser Werte angegeben wird, gelten alle Werte als innerhalb des Bereichs.
Werte ohne Label im Bereich zulassen
Da lange Stringvariablen keine Wertelabels besitzen, sollte diese Option für solche Variablen immer aktiviert sein.
Nicht-ganzzahlige Werte im Bereich zulassen

Um einen Bereich anzugeben, geben Sie den Minimum- oder Maximumwert oder beide Werte ein. Mit dem Kontrollkästchen können Sie festlegen, dass Werte ohne Label und nichtganzzahlige Werte im Bereich gekennzeichnet werden.

Mit den Steuerelementen zum Festlegen einer Liste können Sie eine Liste gültiger Werte angeben. Werte, die nicht in der Liste befinden, werden als ungültig gekennzeichnet.

Abbildung 2-4
Regeln für eine Variable: Listendefinition

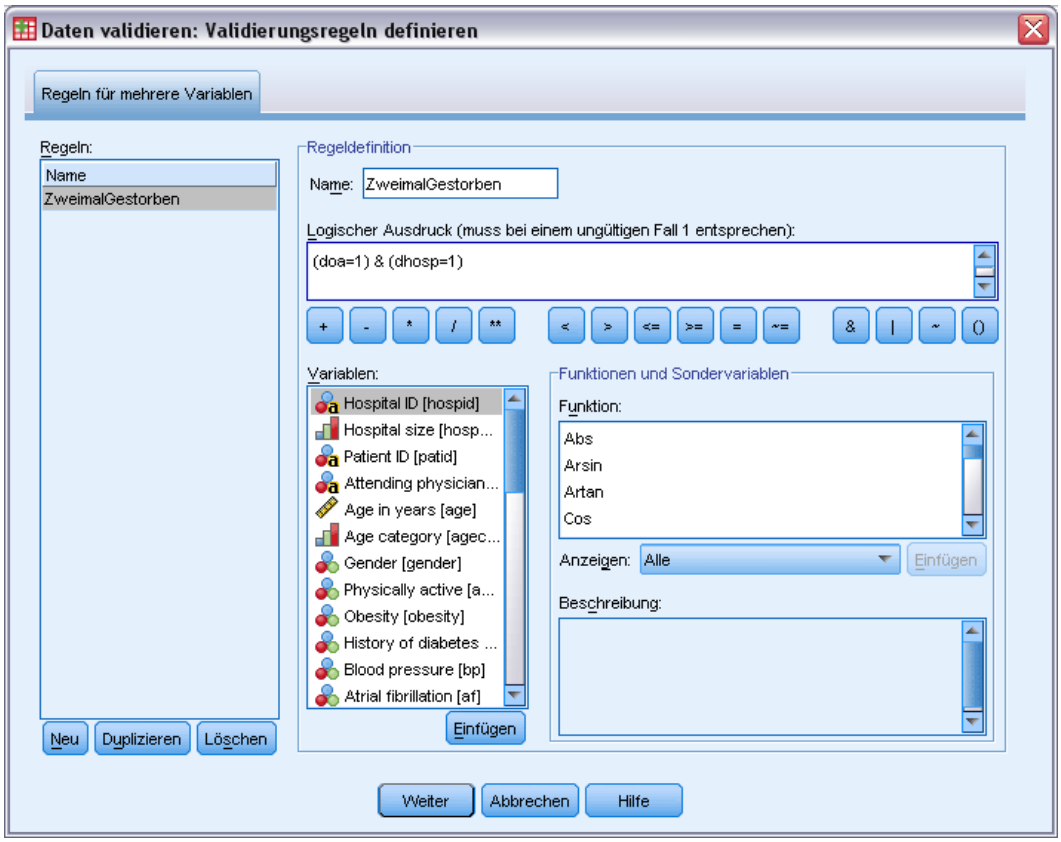


Geben Sie im Gitter die Listenwerte ein. Mit dem Kontrollkästchen legen Sie fest, ob die Groß-/Kleinschreibung berücksichtigt wird, wenn String-Datenwerte gegen die Liste der zulässigen Werte geprüft werden.

- **Benutzerdefinierte fehlende Werte zulassen.** Hiermit wird festgelegt, ob benutzerdefinierte fehlende Werte als ungültig gekennzeichnet werden.
- **Systemdefinierte fehlende Werte zulassen.** Hiermit wird festgelegt, ob systemdefinierte fehlende Werte als ungültig gekennzeichnet werden. Dies gilt nicht für Regeln für Strings.
- **Leere Werte zulassen.** Hiermit wird festgelegt, ob leere String-Werte als ungültig gekennzeichnet werden. Dies gilt nur für Regeln für Strings.

Definieren von Regeln für mehrere Variablen

Abbildung 2-5
Dialogfeld "Validierungsregeln definieren", Registerkarte "Regeln für mehrere Variablen"



Auf der Registerkarte "Regeln für mehrere Variablen" können Sie Validierungsregeln für mehrere Variablen erstellen, anzeigen lassen und ändern.

Regeln. Die Liste enthält die Validierungsregeln für mehrere Variablen nach Namen. Wenn Sie das Dialogfeld öffnen, wird eine Platzhalter-Regel mit dem Namen "MehrVarRegel 1" angezeigt. Unter der Liste "Regeln" werden folgende Schaltflächen angezeigt:

- **Neu.** Fügt einen neuen Eintrag am Ende der Liste "Regeln" hinzu. Die Regel wird ausgewählt und erhält den Namen "MehrVarRegel *n*". Hierbei ist *n* eine Ganzzahl, sodass der Name der Regel unter den Regeln für eine oder mehrere Variablen eindeutig ist.

- **Duplizieren.** Fügt eine Kopie der ausgewählten Regel am Ende der Liste “Regeln” hinzu. Der Name der Regel wird so angepasst, dass er unter den Regeln für eine oder mehrere Variablen eindeutig ist. Wenn Sie beispielsweise “MehrVarRegel 1” duplizieren, erhält die erste duplizierte Regel den Namen “Kopie von MehrVarRegel 1”, die zweite den Namen “Kopie (2) von MehrVarRegel 1” usw.
- **Löschen.** Löscht die ausgewählte Regel.

Regeldefinition. Mit diesen Steuerelementen können Sie die Eigenschaften für eine ausgewählte Regel anzeigen lassen und festlegen.

- **Name.** Der Name der Regel muss unter den Regeln für eine oder mehrere Variablen eindeutig sein.
- **Logischer Ausdruck.** Im Wesentlichen ist dies die Regeldefinition. Die Auswertung des Ausdrucks für einen ungültigen Fall muss 1 entsprechen.

Erstellen von Ausdrücken

▶ Um einen Ausdruck zu erstellen, fügen Sie die Komponenten in das Feld “Logischer Ausdruck” ein oder geben den Ausdruck direkt in dieses Feld ein.

- Sie können Funktionen oder häufig verwendete Systemvariablen einfügen, indem Sie eine Gruppe aus der Liste “Funktion” auswählen und in der Liste “Funktionen und Sodervariablen” auf die Funktion bzw. Variable doppelklicken (oder die Funktion bzw. Variable auswählen und auf Einfügen klicken). Geben Sie alle durch Fragezeichen gekennzeichneten Parameter an (gilt nur für Funktionen). Die Funktionsgruppe mit der Beschriftung Alle bietet eine Liste aller verfügbaren Funktionen und Systemvariablen. Eine kurze Beschreibung der aktuell ausgewählten Funktion oder Variablen wird in einem speziellen Bereich des Dialogfelds angezeigt.
- String-Konstanten müssen in Anführungszeichen oder Apostrophe eingeschlossen werden.
- Wenn die Werte Dezimalstellen enthalten, muss ein Punkt (.) als Dezimaltrennzeichen verwendet werden.

Kapitel
3

Daten validieren

Im Dialogfeld “Daten validieren” können Sie verdächtige oder ungültige Fälle, Variablen und Datenwerte in der Arbeitsdatei identifizieren.

Beispiel. Eine Datenanalytikerin muss für ihren Auftraggeber einen monatlichen Bericht über die Kundenzufriedenheit zusammenstellen. Die monatlich erhaltenen Daten müssen einer Qualitätsprüfung unterzogen werden. Dabei muss nach ungültigen Kunden-IDs, Variablenwerten außerhalb des Bereichs sowie Kombinationen von Variablenwerten gesucht werden, die häufig fehlerhaft eingegeben werden. Im Dialogfeld “Daten validieren” kann die Analytikerin die Variablen angeben, durch die Kunden eindeutig identifiziert werden, Regeln für gültigen Wertebereiche einzelner Variablen definieren und Regeln zum Erkennen unmöglicher Kombinationen für mehrere Variablen definieren. Die Prozedur liefert einen Bericht der Problemfälle und -variablen. Darüber hinaus weisen die Daten in jedem Monat die gleichen Datenelemente auf, sodass die Analytikerin in der Lage ist, die Regeln im folgenden Monat auf die neue Datendatei anzuwenden.

Statistiken. Die Prozedur erzeugt Listen von Variablen, Fällen und Datenwerten, die verschiedene Prüfungen nicht bestehen, Häufigkeiten der Verletzung von Regeln für einzelne oder mehrere Variablen sowie einfache deskriptive Auswertungen der Analysevariablen.

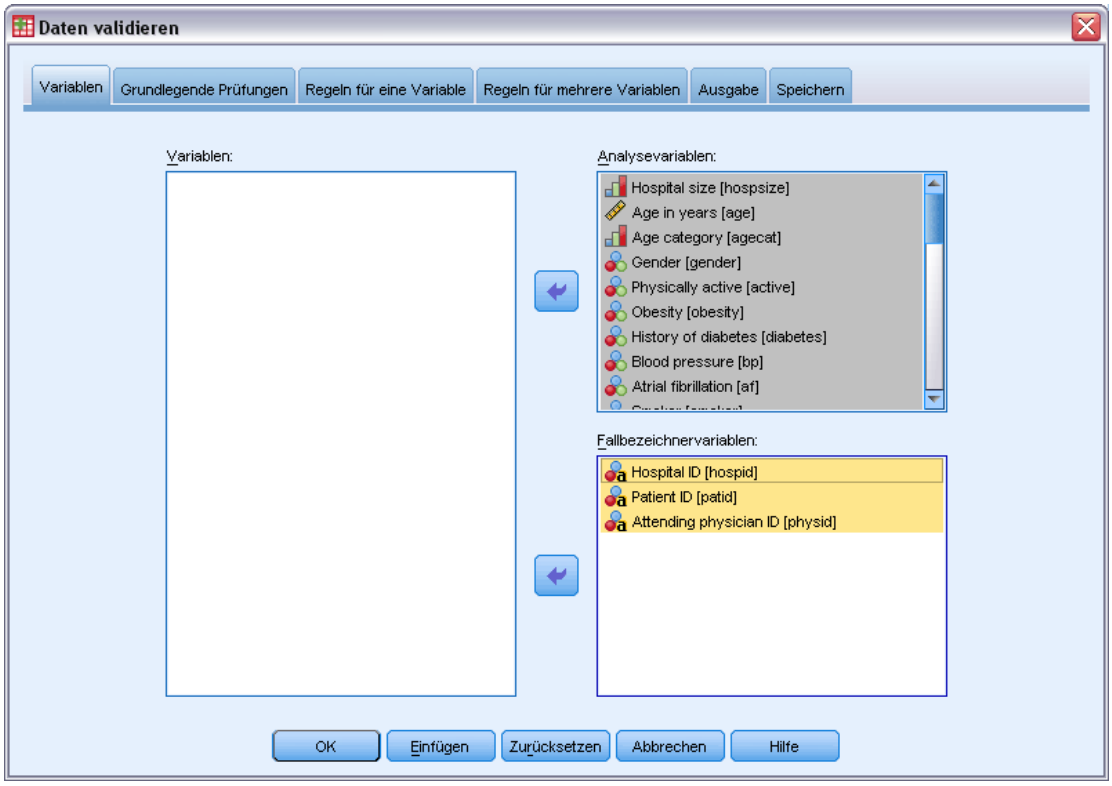
Gewichtungen. Die Prozedur ignoriert Angaben zur Gewichtungsvariablen und behandelt diese stattdessen wie jede andere Analysevariable.

So validieren Sie Daten:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Daten
  Validierung
    Daten validieren...

Abbildung 3-1
Dialogfeld "Daten validieren", Registerkarte "Variablen"



- Wählen Sie eine oder mehrere Analysevariablen aus, die durch grundlegende Variablenprüfungen oder Validierungsregeln für eine Variable validiert werden sollen.

Sie haben außerdem folgende Möglichkeiten:

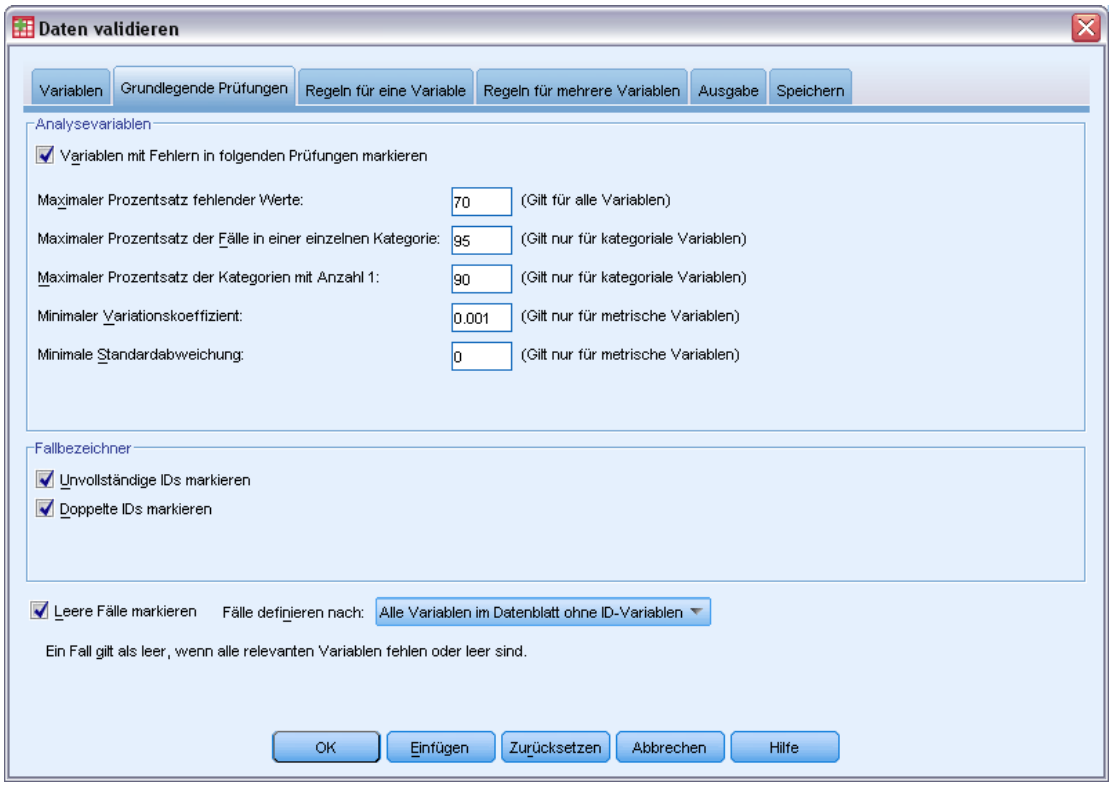
- Klicken Sie auf die Registerkarte Regeln für mehrere Variablen, und wenden Sie eine oder mehrere Regeln für mehrere Variablen an.

Die folgenden Optionen sind verfügbar:

- Wählen Sie eine oder mehrere Fallbezeichnervariablen aus, um nach doppelten oder unvollständigen IDs zu suchen. Fallbezeichnervariablen werden auch zum Beschriften der fallweisen Ausgabe verwendet. Wenn mehr als eine Fallbezeichnervariable angegeben wurde, wird die Kombination der Werte als Fallbezeichner behandelt.

Daten validieren: Grundlegende Prüfungen

Abbildung 3-2
Dialogfeld "Daten validieren", Registerkarte "Grundlegende Prüfungen"



Auf der Registerkarte "Grundlegende Prüfungen" können Sie grundlegende Prüfverfahren für Analysevariablen, Fallbezeichner und ganze Fälle auswählen.

Analysevariablen. Wenn Sie auf der Registerkarte "Variablen" Analysevariablen ausgewählt haben, können Sie die folgenden Gültigkeitsprüfungen auswählen. Mit den Kontrollkästchen können Sie die einzelnen Prüfungen aktivieren oder deaktivieren.

- **Maximaler Prozentsatz fehlender Werte.** Gibt Analysevariablen aus, bei denen der prozentuale Anteil fehlender Werte den angegebenen Wert übersteigt. Der angegebene Wert muss eine positive Zahl kleiner oder gleich 100 sein.
- **Maximaler Prozentsatz der Fälle in einer einzelnen Kategorie.** Wenn kategoriale Analysevariablen vorhanden sind, werden bei dieser Option kategoriale Analysevariablen ausgegeben, bei denen der prozentuale Anteil der Fälle, die eine einzelne nichtfehlende Kategorie darstellen, den angegebenen Wert übersteigt. Der angegebene Wert muss eine positive Zahl kleiner oder gleich 100 sein. Der Prozentsatz entspricht dem Anteil der Fälle mit nichtfehlenden Werten der Variablen.
- **Maximaler Prozentsatz der Kategorien mit Anzahl 1.** Wenn kategoriale Analysevariablen vorhanden sind, werden bei dieser Option kategoriale Analysevariablen ausgegeben, bei denen der prozentuale Anteil der Kategorien der Variablen, die nur einen Fall enthalten,

den angegebenen Wert übersteigt. Der angegebene Wert muss eine positive Zahl kleiner oder gleich 100 sein.

- **Minimaler Variationskoeffizient.** Wenn metrische Analysevariablen vorhanden sind, werden bei dieser Option metrische Analysevariablen ausgegeben, bei denen der absolute Wert des Variationskoeffizienten kleiner als der angegebene Wert ist. Diese Option betrifft nur Variablen mit einem von 0 abweichenden Mittelwert. Der angegebene Wert muss eine nichtnegative Zahl sein. Durch Angabe von 0 wird die Prüfung des Variationskoeffizienten deaktiviert.
- **Minimale Standardabweichung.** Wenn metrische Analysevariablen vorhanden sind, werden bei dieser Option metrische Analysevariablen ausgegeben, deren Standardabweichung kleiner als der angegebene Wert ist. Der angegebene Wert muss eine nichtnegative Zahl sein. Durch Angabe von 0 wird die Prüfung der Standardabweichung deaktiviert.

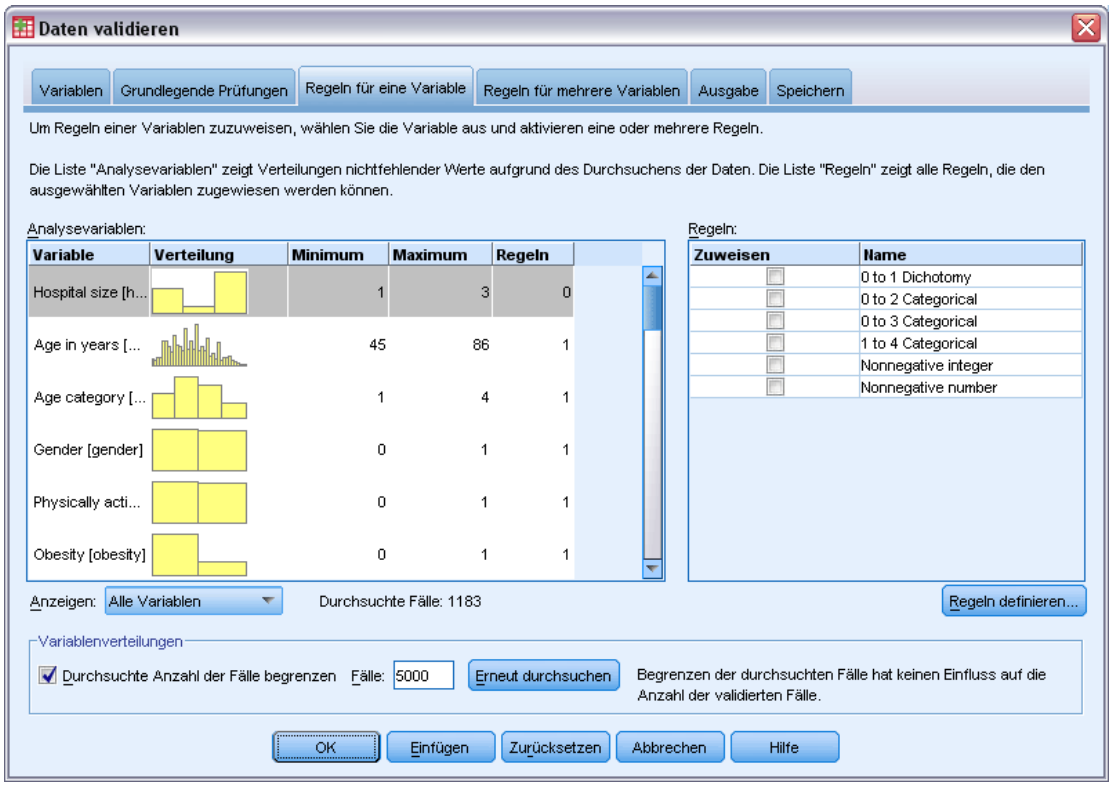
Fallbezeichner. Wenn Sie auf der Registerkarte “Variablen” Fallbezeichnervariablen ausgewählt haben, können Sie die folgenden Gültigkeitsprüfungen auswählen.

- **Unvollständige IDs markieren.** Bei dieser Option werden Fälle mit unvollständigen Fallbezeichnern ausgegeben. Ein Bezeichner wird bei einem gegebenen Fall als unvollständig betrachtet, wenn der Wert einer ID-Variable leer ist oder fehlt.
- **Doppelte IDs markieren.** Bei dieser Option werden Fälle mit doppelten Fallbezeichnern ausgegeben. Unvollständige Fallbezeichner werden aus der Menge der möglichen doppelten Werte ausgeschlossen.

Leere Fälle markieren. Bei dieser werden Fälle ausgegeben, bei denen alle Variablen leer sind oder fehlen. Sie können festlegen, ob zum Identifizieren leerer Fälle alle Variablen in der Datei (mit Ausnahme von ID-Variablen) oder nur die auf der Registerkarte “Variablen” ausgewählten Analysevariablen herangezogen werden sollen.

Daten validieren: Regeln für eine Variable

Abbildung 3-3
Dialogfeld "Daten validieren", Registerkarte "Regeln für eine Variable"



Auf der Registerkarte "Regeln für eine Variable" werden verfügbare Validierungsregeln für eine Variable angezeigt, die Sie auf die Analysevariablen anwenden können. Um weitere Regeln für einzelne Variablen zu definieren, klicken Sie auf Regeln definieren. Für weitere Informationen siehe Definieren von Regeln für eine Variable in Kapitel 2 auf S. 4.

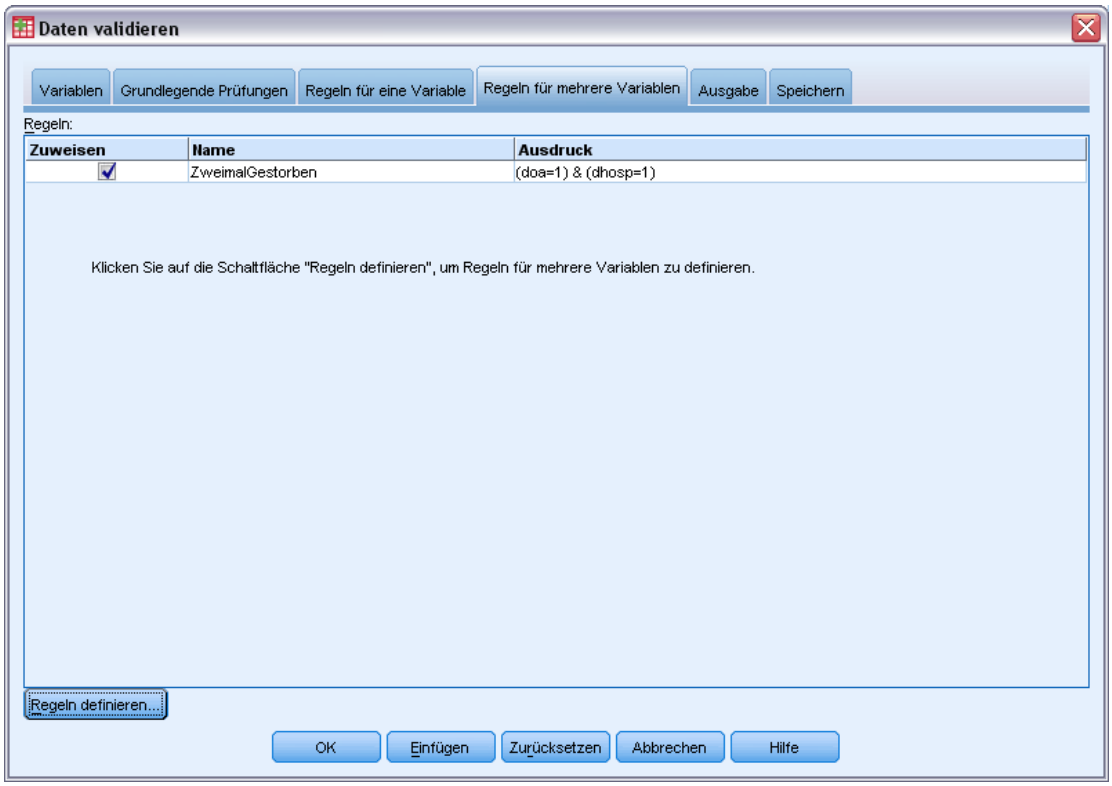
Analysevariablen. In der Liste werden Analysevariablen aufgeführt, ihre Verteilungen zusammengefasst und die Anzahl der Regeln angezeigt, die auf jede Variable angewendet werden. Beachten Sie, dass benutzerdefinierte und systemdefinierte fehlende Werte nicht in den Zusammenfassungen enthalten sind. Durch die Dropdown-Liste "Anzeige" wird gesteuert, welche Variablen angezeigt werden. Zur Auswahl stehen Alle Variablen, Numerische Variablen, String-Variablen und Datumsvariablen.

Regeln. Um Regeln auf Analysevariablen anzuwenden, wählen Sie eine oder mehrere Variablen aus, und aktivieren Sie in der Liste "Regeln" alle anzuwendenden Regeln. In der Liste "Regeln" werden nur Regeln aufgeführt, die für die ausgewählten Analysevariablen geeignet sind. Wenn beispielsweise numerische Variablen ausgewählt wurden, werden nur numerische Regeln angezeigt. Wurde eine String-Variable ausgewählt, werden nur String-Regeln angezeigt. Wenn keine Analysevariablen ausgewählt wurden oder die ausgewählten Variablen unterschiedliche Datentypen aufweisen, werden keine Regeln angezeigt.

Variablenverteilungen. Die in der Liste “Analysevariablen” angezeigten Verteilungszusammenfassungen können auf allen Fällen beruhen oder auf einer Durchsuchung der ersten *n* Fälle. Dies wird im Textfeld “Fälle” festgelegt. Durch Klicken auf Erneut durchsuchen werden die Verteilungszusammenfassungen aktualisiert.

Daten validieren: Regeln für mehrere Variablen

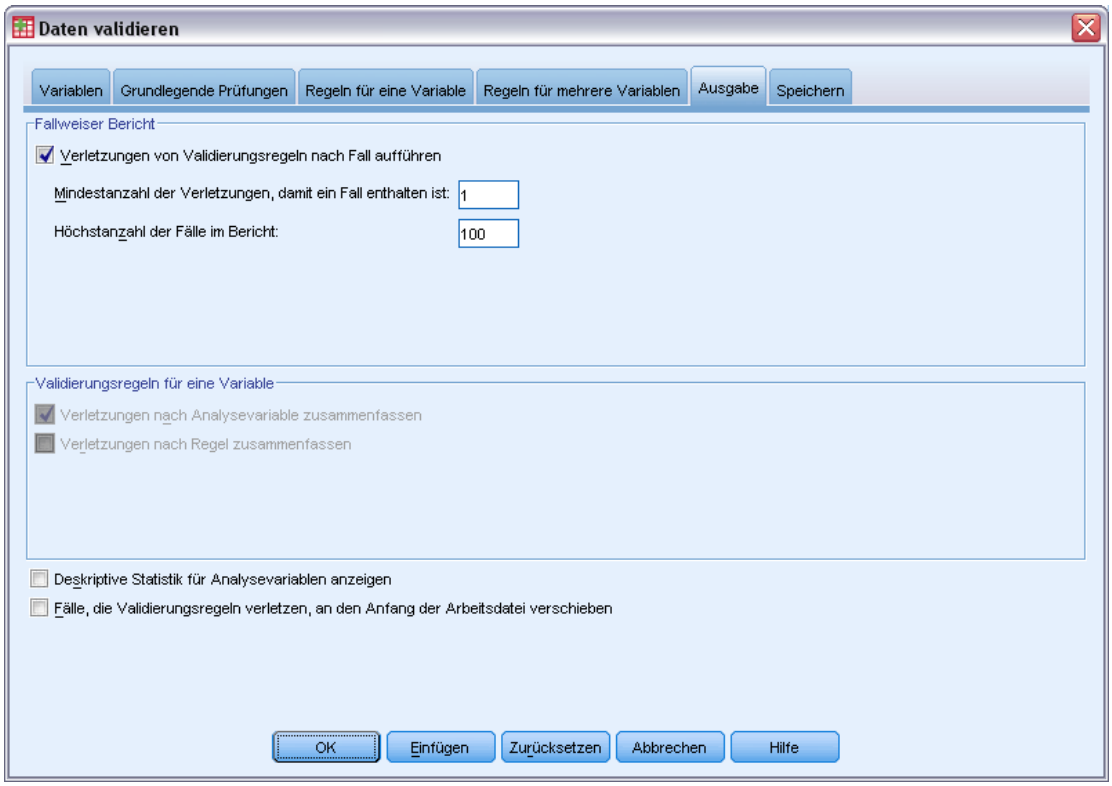
Abbildung 3-4
Dialogfeld “Daten validieren,” Registerkarte “Regeln für mehrere Variablen”



Auf der Registerkarte “Regeln für mehrere Variablen” werden verfügbare Regeln für mehrere Variablen angezeigt, die Sie auf die Daten anwenden können. Um weitere Regeln für mehrere Variablen zu definieren, klicken Sie auf Regeln definieren. Für weitere Informationen siehe Definieren von Regeln für mehrere Variablen in Kapitel 2 auf S. 6.

Daten validieren: Ausgabe

Abbildung 3-5
Dialogfeld "Daten validieren", Registerkarte "Ausgabe"



Fallweiser Bericht. Wenn Sie Validierungsregeln für eine oder mehrere Variablen ausgewählt haben, können Sie einen Bericht anfordern, der die Verletzungen der Validierungsregeln für einzelne Fälle enthält.

- **Mindestanzahl der Verletzungen, damit ein Fall enthalten ist.** Mit dieser Option wird die Mindestanzahl der Verletzungen angegeben, die erforderlich sind, damit ein Fall in den Bericht aufgenommen wird. Geben Sie eine positive Ganzzahl ein.
- **Höchstanzahl der Fälle im Bericht.** Mit dieser Option wird die Höchstanzahl der Fälle angegeben, die im Fallbericht enthalten sein soll. Geben Sie eine positive ganze Zahl kleiner oder gleich 1000 ein.

Validierungsregeln für eine Variable. Wenn Sie Validierungsregeln für einzelne Variablen angewendet haben, können Sie auswählen, ob und wie die Ergebnisse angezeigt werden sollen.

- **Verletzungen nach Analysevariable zusammenfassen.** Bei dieser Option werden für jede Analysevariable alle Validierungsregeln für eine Variable aufgeführt, die verletzt wurden, und die Anzahl der Werte angegeben, die eine Verletzung der einzelnen Regeln darstellen.

Außerdem wird für jede Variable die Gesamtanzahl der Verletzungen von Regeln für eine Variable ausgegeben.

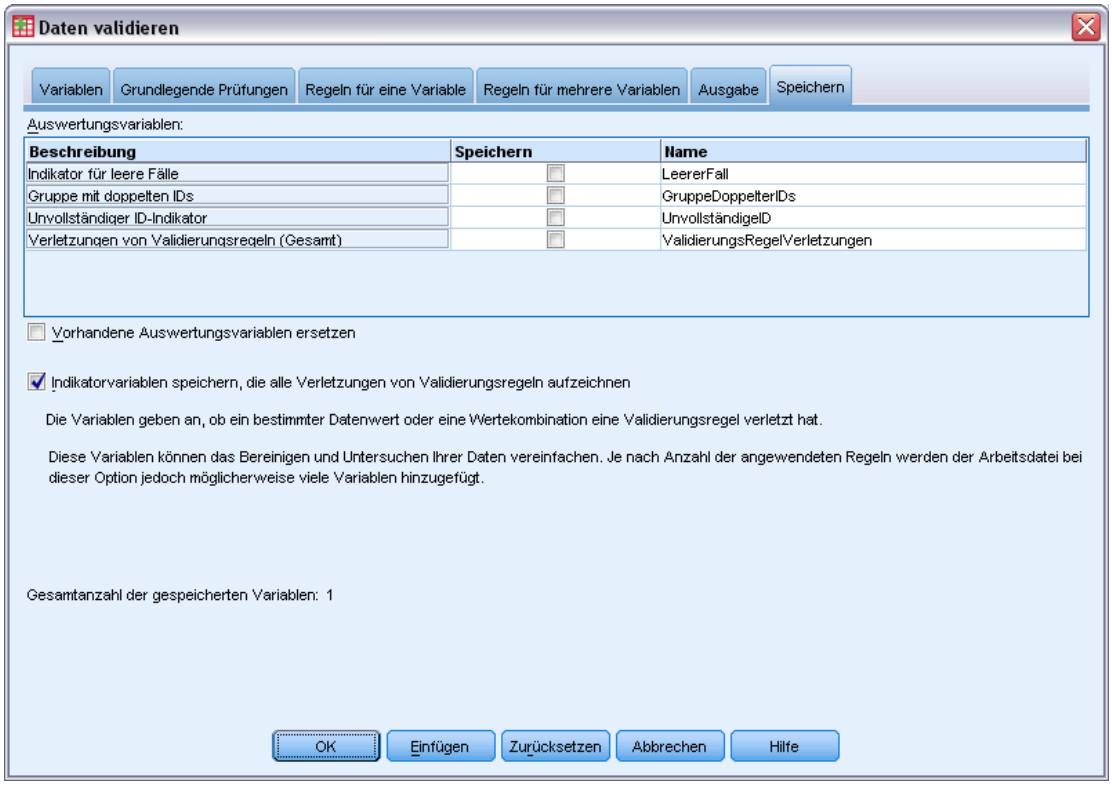
- **Verletzungen nach Regel zusammenfassen.** Bei dieser Option werden für jede Validierungsregel für eine Variable die Variablen ausgegeben, die die Regeln verletzen, und die Anzahl der ungültigen Werte pro Variable angegeben. Außerdem wird variablenübergreifend die Gesamtanzahl der Werte ausgegeben, die eine Verletzung der einzelnen Regeln darstellen.

Deskriptive Statistik für Analysevariablen anzeigen. Mit dieser Option können Sie deskriptive Statistiken für Analysevariablen anfordern. Für jede kategoriale Variable wird eine Häufigkeitstabelle erzeugt. Für metrische Variablen wird eine Tabelle mit Auswertungsstatistiken erzeugt, darunter der Mittelwert, die Standardabweichung, das Minimum und das Maximum.

Fälle, die Validierungsregeln verletzen, an den Anfang der Arbeitsdatei verschieben. Bei dieser Option werden Fälle mit Verletzungen von Regeln für eine oder mehrere Variablen an den Anfang der ARbeitsdatei verschoben, damit sie einfacher aufgefunden werden können.

Daten validieren: Speichern

Abbildung 3-6
Dialogfeld "Daten validieren", Registerkarte "Speichern"



Mithilfe der Registerkarte "Speichern" können Sie Variablen, bei denen Regelverletzungen verzeichnet wurden, in der Arbeitsdatei speichern.

Auswertungsvariablen. Hierbei handelt es sich um einzelne Variablen, die gespeichert werden können. Aktivieren Sie die Kontrollkästchen der zu speichernden Variablen. Für die Variablen sind Standardnamen vorgegeben, die Sie bearbeiten können.

- **Indikator für leere Fälle.** Leeren Fällen wird der Wert 1 zugeordnet. Alle anderen Fälle werden als 0 codiert. Die Werte der Variablen entsprechen dem Umfang, der auf der Registerkarte "Grundlegende Prüfungen" angegeben wurde.
- **Gruppe mit doppelten IDs.** Fälle, die denselben Fallbezeichner aufweisen (mit Ausnahme von Fällen mit unvollständigen Bezeichnern), erhalten dieselbe Gruppennummer. Fälle mit eindeutigen oder unvollständigen Bezeichnern werden als 0 codiert.
- **Unvollständiger ID-Indikator.** Fälle mit leeren oder unvollständigen Fallbezeichnern erhalten den Wert 1. Alle anderen Fälle werden als 0 codiert.
- **Verletzungen von Validierungsregeln.** Dies ist die Gesamtanzahl der Verletzungen von Validierungsregeln für eine oder mehrere Variablen pro Fall.

Vorhandene Auswertungsvariablen ersetzen. In der Datendatei gespeicherte Variablen müssen eindeutige Namen aufweisen. Wenn dies nicht der Fall ist, werden Variablen mit demselben Namen ersetzt.

Indikatorvariablen speichern, die alle Verletzungen von Validierungsregeln aufzeichnen. Bei dieser Option wird ein vollständiger Bericht über die Verletzungen der Validierungsregeln gespeichert. Jede Variable entspricht der Anwendung einer Validierungsregel und weist den Wert 1 auf, wenn der Fall die Regel verletzt, oder den Wert 0, wenn die Regel nicht verletzt wird.

Kapitel

4

Automatisierte Datenaufbereitung

Die Aufbereitung von Daten zur Analyse ist einer der wichtigsten Schritte in jedem Projekt – und gewöhnlich auch einer der zeitaufwendigsten. Die automatisierte Datenaufbereitung (ADP) übernimmt diese Aufgabe für Sie. Sie analysiert Ihre Daten und identifiziert Problemlösungen, findet problematische oder wahrscheinlich nicht nützliche Felder, leitet zum passenden Zeitpunkt neue Attribute ab und verbessert die Leistungsfähigkeit durch intelligente Screening-Methoden. Sie können den Algorithmus **vollautomatisch** verwenden und so Problemlösungen auswählen und anwenden oder Sie können ihn **interaktiv** verwenden und so die Änderungen in einer Vorschau betrachten, bevor sie vorgenommen werden, und sie gegebenenfalls akzeptieren oder ablehnen.

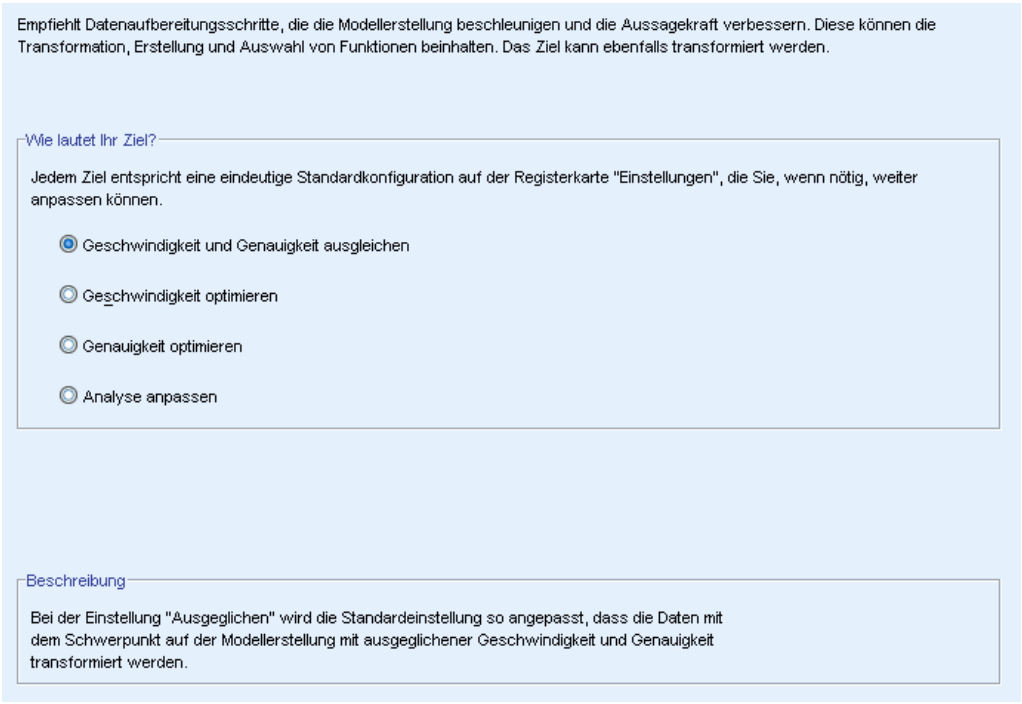
Mit ADP können Sie Ihre Daten schnell und einfach für die Modellerstellung aufbereiten, ohne über Vorkenntnisse der dazugehörigen statistischen Konzepte verfügen zu müssen. Modelle lassen sich damit schneller erstellen und scoren; zudem verbessert sich mit ADP die Robustheit automatisierter Modellierungsprozesse.

Anmerkung: Wenn die ADP ein Feld für die Analyse vorbereitet, erstellt sie ein neues Feld, das die Anpassungen oder Transformationen enthält, anstatt die bestehenden Werte und Eigenschaften des alten Felds zu ersetzen. Das alte Feld wird bei der weiteren Analyse nicht verwendet; seine Rolle wird auf “Keine” gesetzt.

Beispiel. Eine Versicherungsgesellschaft mit beschränkten Ressourcen für die Untersuchung der Versicherungsansprüche von Hauseigentümern möchte ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen. Vor Erstellung des Modells bereiten sie die Daten für die Modellierung mithilfe der automatisierten Datenaufbereitung vor. Da sie die vorgeschlagenen Transformationen zunächst überprüfen möchten, bevor die Transformationen angewendet werden, nutzen sie die automatisierte Datenaufbereitung im interaktiven Modus. Für weitere Informationen siehe Interaktive Verwendung der automatisierten Datenaufbereitung in Kapitel 8 auf S. 82.

Eine Gruppe in der Kraftfahrzeugindustrie erfasst die Verkaufszahlen verschiedener Personenkraftwagen. Um starke und schwache Modelle identifizieren zu können, soll eine Beziehung zwischen den Fahrzeugverkaufszahlen und den Fahrzeugeigenschaften hergestellt werden. Zur Vorbereitung der Daten für die Analyse wird die automatisierte Datenaufbereitung verwendet. Es werden Modelle mit Daten “vor” und “nach” der Aufbereitung erstellt, um zu sehen, wie sich die Ergebnisse unterscheiden. Für weitere Informationen siehe Automatische Verwendung der automatisierten Datenaufbereitung in Kapitel 8 auf S. 95.

Abbildung 4-1
Registerkarte "Ziel" in der automatisierten Datenaufbereitung


Empfiehlt Datenaufbereitungsschritte, die die Modellerstellung beschleunigen und die Aussagekraft verbessern. Diese können die Transformation, Erstellung und Auswahl von Funktionen beinhalten. Das Ziel kann ebenfalls transformiert werden.
Wie lautet Ihr Ziel?
Jedem Ziel entspricht eine eindeutige Standardkonfiguration auf der Registerkarte "Einstellungen", die Sie, wenn nötig, weiter anpassen können.
Geschwindigkeit und Genauigkeit ausgleichen
Geschwindigkeit optimieren
Genauigkeit optimieren
Analyse anpassen
Beschreibung
Bei der Einstellung "Ausgeglichen" wird die Standardeinstellung so angepasst, dass die Daten mit dem Schwerpunkt auf der Modellerstellung mit ausgeglichener Geschwindigkeit und Genauigkeit transformiert werden.

Wie lautet Ihr Ziel? Die automatisierte Datenaufbereitung empfiehlt Schritte zur Datenaufbereitung, die sich auf die Geschwindigkeit auswirken, mit der andere Algorithmen Modelle erstellen können und die Vorhersagekraft dieser Modelle verbessern. Diese können die Transformation, Erstellung und Auswahl von Funktionen beinhalten. Das Ziel kann ebenfalls transformiert werden. Sie können die Prioritäten der Modellerstellung festlegen, auf die sich die Datenaufbereitung konzentrieren sollte.

- **Geschwindigkeit und Genauigkeit ausgleichen.** Diese Option bereitet die Daten auf und sorgt dabei für eine ausgeglichene Priorität zwischen der Geschwindigkeit, mit der Daten durch die Modellerstellung verarbeitet werden, und der Genauigkeit der Vorhersagen.
- **Geschwindigkeit optimieren.** Diese Option bereitet die Daten auf und gibt dabei der Geschwindigkeit Vorrang, mit der Daten durch Modellerstellungsalgorithmen verarbeitet werden. Wählen Sie diese Option, wenn Sie mit sehr großen Daten-Sets arbeiten oder nach einer schnellen Antwort suchen.
- **Genauigkeit optimieren.** Diese Option bereitet die Daten auf und gibt dabei der Genauigkeit der durch Modellerstellungsalgorithmen erzeugten Vorhersagen Vorrang.
- **Analyse anpassen** Wählen Sie diese Option, wenn Sie den Algorithmus auf der Registerkarte "Einstellungen" manuell ändern wollen. Beachten Sie, dass diese Einstellung automatisch ausgewählt wird, wenn Sie anschließend Änderungen auf der Registerkarte "Einstellungen" vornehmen, die mit einem der anderen Ziele nicht kompatibel sind.

So rufen Sie die automatische Datenaufbereitung ab

Wählen Sie die folgenden Befehle aus den Menüs aus:

Transformieren
  Daten für Modellierung vorbereiten
    Automatisch...

▶ Klicken Sie auf Ausführen.

Die folgenden Optionen sind verfügbar:

- Geben Sie ein Ziel auf der Registerkarte “Ziel” an.
- Geben Sie Feldzuweisungen auf der Registerkarte “Felder” an.
- Geben Sie Experteneinstellungen auf der Registerkarte “Einstellungen” an.

So rufen Sie die interaktive Datenaufbereitung ab

Wählen Sie die folgenden Befehle aus den Menüs aus:

Transformieren
  Daten für Modellierung vorbereiten
    Interaktiv...

▶ Klicken Sie auf Analysieren in der Symbolleiste im oberen Bereich des Dialogfelds.

▶ Klicken Sie auf die Registerkarte “Analyse” und überprüfen Sie die folgenden Schritte der Datenaufbereitung.

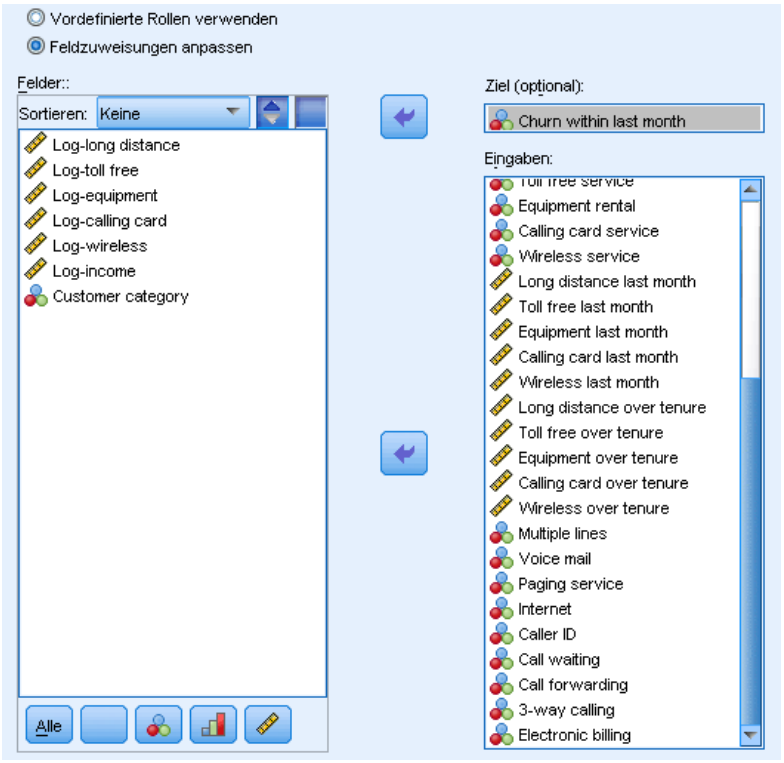
▶ Sind alle Angaben korrekt, klicken Sie auf Ausführen. Wenn nicht, klicken Sie auf Analyse löschen, ändern die Einstellungen nach Ihren Wünschen und klicken dann auf Analysieren.

Die folgenden Optionen sind verfügbar:

- Geben Sie ein Ziel auf der Registerkarte “Ziel” an.
- Geben Sie Feldzuweisungen auf der Registerkarte “Felder” an.
- Geben Sie Experteneinstellungen auf der Registerkarte “Einstellungen” an.
- Speichern Sie die vorgeschlagenen Schritte der Datenaufbereitung in eine XML-Datei mit einem Klick auf XML speichern.

Registerkarte "Felder"

Abbildung 4-2
Registerkarte "Felder" in der automatisierten Datenaufbereitung



Die Registerkarte "Felder" gibt an, welche Felder zur weiteren Analyse aufbereitet werden sollen.

Vordefinierte Rollen verwenden Diese Option greift auf bestehende Feldinformationen zurück. Wenn ein einzelnes Feld mit einer Rolle als "Ziel" vorhanden ist, wird es als Ziel verwendet; in allen anderen Fällen ist kein Ziel vorhanden. Alle Felder mit der vordefinierten Rolle "Eingabe" werden als Eingaben verwendet. Mindestens ein Eingabefeld ist erforderlich.

Benutzerdefinierte Feldzuweisungen verwenden Wenn Sie Feldrollen durch Verschieben von Feldern aus ihren Standardlisten überschreiben, springt das Dialogfeld automatisch auf diese Option. Wenn Sie benutzerdefinierte Feldzuweisungen vornehmen, geben Sie die folgenden Felder an:

- **Ziel (optional).** Wählen Sie das Zielfeld aus, wenn Sie Modelle erstellen möchten, für die ein Ziel erforderlich ist. Dies gleicht in etwa der Einstellung der Feldrolle auf "Ziel".
- **Eingaben.** Wählen Sie mindestens ein Eingabefeld aus. Dies gleicht in etwa der Einstellung der Feldrolle auf "Eingabe".

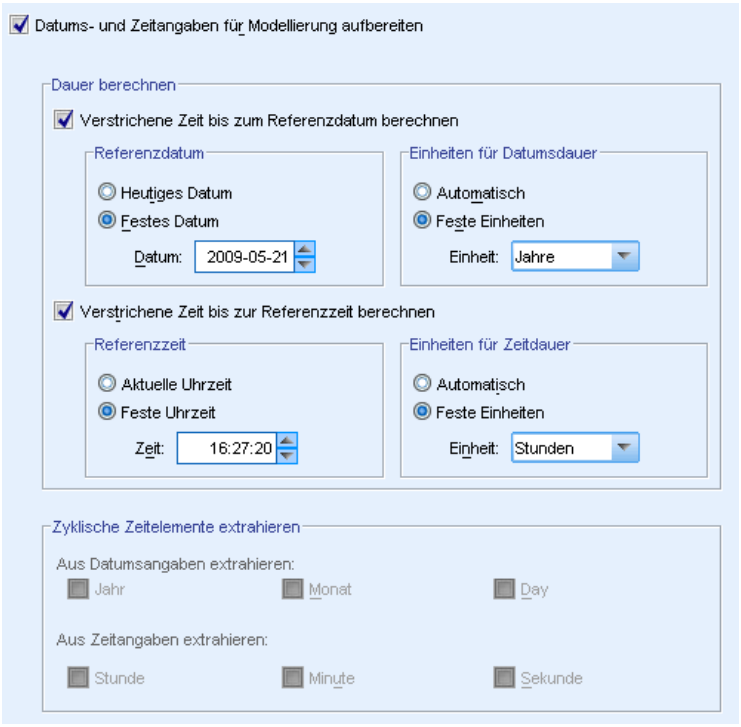
Registerkarte "Einstellungen"

Die Registerkarte "Einstellungen" enthält mehrere unterschiedliche Gruppen von Einstellungen, die Sie ändern können, um genau festzulegen, wie der Algorithmus Ihre Daten verarbeiten soll. Wenn Sie an den Standardeinstellungen Änderungen vornehmen, die mit den anderen Zielen nicht

kompatibel sind, wird auf der Registerkarte “Ziel” automatisch die Option Analyse anpassen ausgewählt.

Datum und Uhrzeit aufbereiten

Abbildung 4-3
Automatisierte Datenaufbereitung – Datum und Uhrzeit aufbereiten – Einstellungen



Viele Modellierungsalgorithmen sind nicht in der Lage, Datums- und Zeitangaben direkt zu behandeln; mit diesen Einstellungen können Sie neue Laufzeitdaten ableiten, die Sie in Ihren bestehenden Daten als Modelleingaben aus Datums- und Zeitangaben verwenden können. Die Felder mit Datums- und Zeitangaben müssen mit Datums- oder Zeitspeichertypen vordefiniert sein. Die ursprünglichen Datums- und Zeitfelder werden nicht als Modelleingaben nach der automatisierten Datenaufbereitung empfohlen.

Datums- und Zeitangaben für Modellierung aufbereiten. Durch Deaktivieren dieser Option werden alle anderen Datums- und Zeiteingaben deaktiviert und die Auswahl beibehalten.

Verstrichene Zeit bis zum Referenzdatum berechnen. Errechnet die Anzahl der Jahre/Monate/Tage seit einem Referenzdatum für jede Variable, die Datumsangaben enthält.

- **Referenzdatum.** Geben Sie das Datum an, ab dem die Dauer bezüglich der Datumsinformationen in den Eingabedaten berechnet wird. Durch die Auswahl von Heutiges Datum wird das aktuelle Systemdatum stets verwendet, wenn ADP ausgeführt wird. Um ein bestimmtes Datum zu verwenden, wählen Sie Festes Datum und geben Sie das erforderliche Datum ein.
- **Einheiten für Datumsdauer.** Legen Sie fest, ob ADP die Einheit der Datumsdauer automatisch bestimmen soll, oder wählen Sie Feste Einheiten für Jahre, Monate oder Tage.

Verstrichene Zeit bis zur Referenzzeit berechnen. Errechnet die Anzahl der Stunden/Minuten/Sekunden seit einer Referenzzeit für jede Variable, die Uhrzeiten enthält.

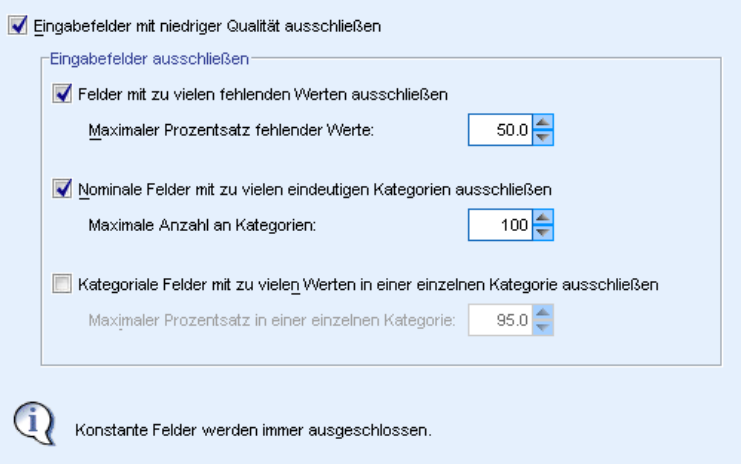
- **Referenzzeit.** Geben Sie die Zeit an, ab der die Dauer bezüglich der Zeitinformationen in den Eingabedaten berechnet wird. Durch die Auswahl von Aktuelle Uhrzeit wird die aktuelle Systemzeit stets verwendet, wenn ADP ausgeführt wird. Um eine bestimmte Uhrzeit zu verwenden, wählen Sie Feste Uhrzeit und geben Sie die erforderlichen Daten ein.
- **Einheiten für Zeitdauer.** Legen Sie fest, ob ADP die Einheit der Zeitdauer automatisch bestimmen soll, oder wählen Sie Feste Einheiten für Stunden, Minuten oder Sekunden.

Zyklische Zeitelemente extrahieren. Verwenden Sie diese Einstellungen, um ein einzelnes Datums- oder Zeitfeld in ein oder mehrere Felder aufzuteilen. Wenn Sie zum Beispiel alle drei Datumskontrollkästchen auswählen, wird das Eingabedatumsfeld "1954-05-23" in drei Felder aufgeteilt: 1954, 5 und 23, wobei jedes das unter Feldnamen definierte Suffix verwendet und das ursprüngliche Datumsfeld ignoriert wird.

- **Aus Datumsangaben extrahieren.** Legen Sie für eine beliebige Datumseingabe fest, ob Sie Jahre, Monate, Tage oder eine Kombination daraus extrahieren möchten.
- **Aus Zeitangaben extrahieren.** Legen Sie für eine beliebige Zeiteingabe fest, ob Sie Stunden, Minuten, Sekunden oder eine Kombination daraus extrahieren möchten.

Felder ausschließen

Abbildung 4-4
Automatisierte Datenaufbereitung – Felder ausschließen – Einstellungen


Eingabefelder mit niedriger Qualität ausschließen
Eingabefelder ausschließen
Felder mit zu vielen fehlenden Werten ausschließen
Maximaler Prozentsatz fehlender Werte:
50.0
Nominale Felder mit zu vielen eindeutigen Kategorien ausschließen
Maximale Anzahl an Kategorien:
100
Kategoriale Felder mit zu vielen Werten in einer einzelnen Kategorie ausschließen
Maximaler Prozentsatz in einer einzelnen Kategorie:
95.0
Konstante Felder werden immer ausgeschlossen.

Schlechte Datenqualität kann sich negativ auf die Genauigkeit Ihrer Vorhersagen auswirken; Sie können daher die akzeptable Qualitätsstufe für Eingabefunktionen festlegen. Alle konstanten oder 100 % an fehlenden Werten aufweisenden Felder werden automatisch ausgeschlossen.

Eingabefelder mit niedriger Qualität ausschließen. Durch Deaktivieren dieser Option werden alle anderen Befehle "Felder ausschließen" deaktiviert und die Auswahl beibehalten.

Felder mit zu vielen fehlenden Werten ausschließen. Felder mit mehr als dem angegebenen Prozentsatz an fehlenden Werten werden aus der weiteren Analyse ausgeschlossen. Geben Sie einen Wert größer oder gleich 0 ein, was dem Deaktivieren dieser Option entspricht, und

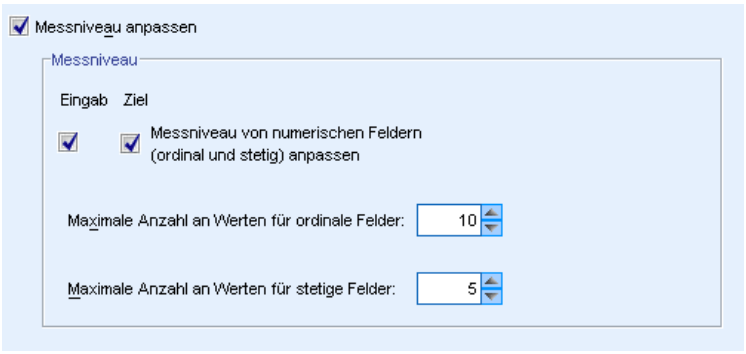
einen Wert kleiner oder gleich 100, so dass die Felder mit allen fehlenden Werten automatisch ausgeschlossen werden. Der Standardwert ist 50.

Nominale Felder mit zu vielen eindeutigen Kategorien ausschließen. Nominale Felder mit mehr als der angegebenen Anzahl an Kategorien werden aus der weiteren Analyse ausgeschlossen. Geben Sie eine positive Ganzzahl ein. Der Standardwert ist 100. Dies ist nützlich für das automatische Entfernen von Feldern aus der Modellierung, die eine datensatzeindeutige Information enthalten, wie zum Beispiel eine ID, eine Adresse oder einen Namen.

Kategoriale Felder mit zu vielen Werten in einer einzelnen Kategorie ausschließen. Ordinale und nominale Felder mit einer Kategorie, die mehr als die angegebene Prozentzahl an Datensätzen enthält, werden aus der weiteren Analyse ausgeschlossen. Geben Sie einen Wert größer oder gleich 0 ein, was dem Deaktivieren dieser Option entspricht, und einen Wert kleiner oder gleich 100, so dass konstante Felder automatisch ausgeschlossen werden. Der Standardwert ist 95.

Messniveau anpassen

Abbildung 4-5
Automatisierte Datenaufbereitung – Messniveau anpassen – Einstellungen



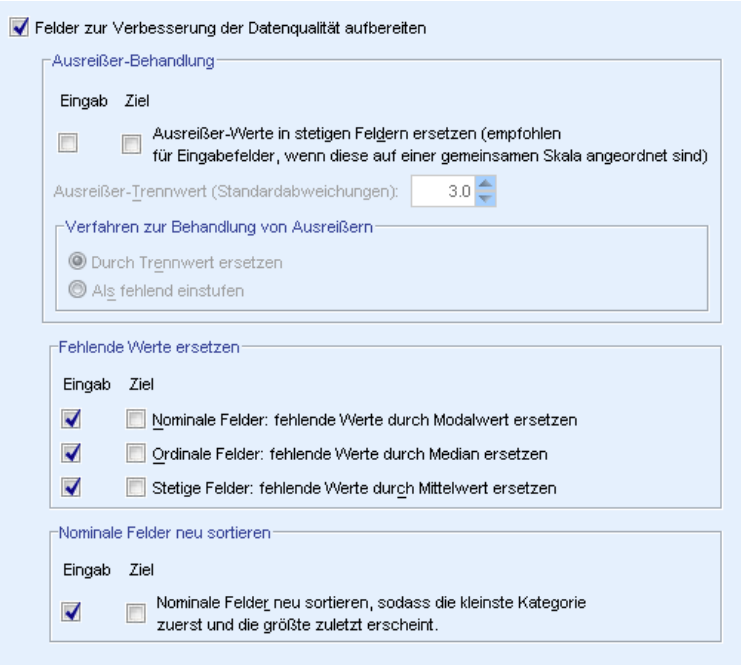
Messniveau anpassen. Durch Deaktivieren dieser Option werden alle anderen Befehle "Messniveau anpassen" deaktiviert und die Auswahl beibehalten.

Messniveau. Legen Sie fest, ob das Messniveau von stetigen Feldern mit "zu wenigen" Werten auf ordinal und von ordinalen Feldern mit "zu vielen" Werten auf stetig angepasst werden kann.

- **Maximale Anzahl an Werten für ordinale Felder.** Ordinale Felder mit mehr als der angegebenen Anzahl an Kategorien werden in stetige Felder umgewandelt. Geben Sie eine positive Ganzzahl ein. Der Standardwert ist 10. Dieser Wert kann größer oder gleich der Mindestanzahl an Werten für stetige Felder sein.
- **Minimale Anzahl an Werten für stetige Felder.** Stetige Felder mit weniger als der angegebenen Anzahl an eindeutigen Werten werden in ordinale Felder umgewandelt. Geben Sie eine positive Ganzzahl ein. Der Standardwert ist 5. Dieser Wert kann kleiner oder gleich der Höchstanzahl an Werten für ordinale Felder sein.

Datenqualität verbessern

Abbildung 4-6
Automatisierte Datenaufbereitung – Datenqualität verbessern – Einstellungen


Felder zur Verbesserung der Datenqualität aufbereiten
Ausreißer-Behandlung
Eingab
Ziel
Ausreißer-Werte in stetigen Feldern ersetzen (empfohlen für Eingabefelder, wenn diese auf einer gemeinsamen Skala angeordnet sind)
Ausreißer-Trennwert (Standardabweichungen):
3.0
Verfahren zur Behandlung von Ausreißern
Durch Trennwert ersetzen
Als fehlend einstufen
Fehlende Werte ersetzen
Eingab
Ziel
Nominale Felder: fehlende Werte durch Modalwert ersetzen
Ordinale Felder: fehlende Werte durch Median ersetzen
Stetige Felder: fehlende Werte durch Mittelwert ersetzen
Nominale Felder neu sortieren
Eingab
Ziel
Nominale Felder neu sortieren, sodass die kleinste Kategorie zuerst und die größte zuletzt erscheint.

Felder zur Verbesserung der Datenqualität aufbereiten. Durch Deaktivieren dieser Option werden alle anderen Einstellungen zu “Datenqualität verbessern” deaktiviert und die Auswahl beibehalten.

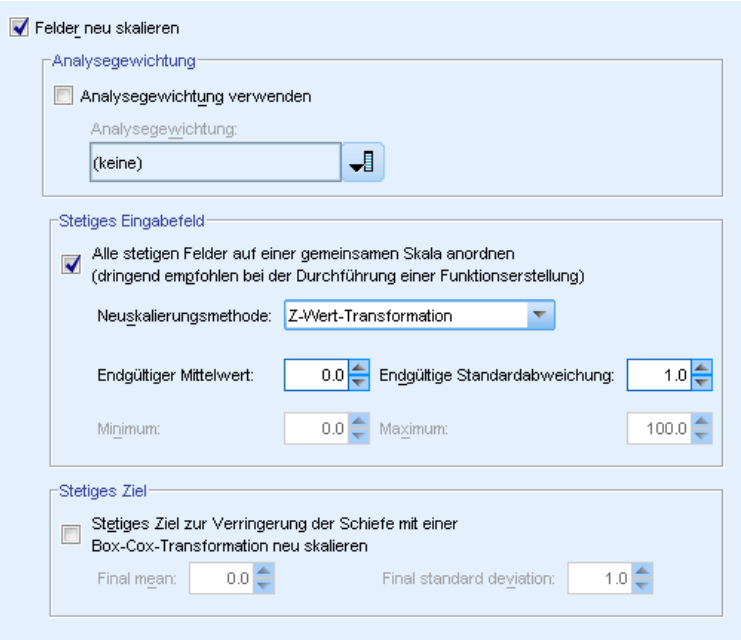
Ausreißer-Behandlung. Legen Sie fest, ob Ausreißer für die Eingaben und Ziele ersetzt werden sollen; wenn ja, geben Sie ein in Standardabweichungen gemessenes Ausreißer-Trennwert-Kriterium und eine Methode zum Ersetzen der Ausreißer an. Ausreißer können entweder durch Entfernen (durch Setzen auf den Trennwert) oder durch Einstufung als fehlende Werte ersetzt werden. Jeder als fehlender Wert eingestufte Ausreißer unterliegt den unten ausgewählten Einstellungen für die Behandlung fehlender Werte.

Fehlende Werte ersetzen. Legen Sie fest, ob fehlende Werte von stetigen, nominalen oder ordinalen Feldern ersetzt werden sollen.

Nominale Felder neu sortieren. Mit dieser Option werden die Werte von nominalen (Set-)Feldern von der kleinsten (am seltensten auftretenden) zur größten (am häufigsten auftretenden) Kategorie umkodiert. Die neuen Feldwerte starten mit 0 als der seltensten Kategorie. Hinweis: Das neue Feld ist numerisch, auch wenn das originale Feld eine Zeichenfolge enthält. Wenn zum Beispiel die Datenwerte eines nominalen Felds “A”, “A”, “A”, “B”, “C”, “C” sind, kodiert die automatisierte Datenaufbereitung “B” zu 0 um, “C” zu 1 und “A” zu 2.

Felder neu skalieren

Abbildung 4-7
Automatisierte Datenaufbereitung – Felder neu skalieren – Einstellungen



Felder neu skalieren. Durch Deaktivieren dieser Option werden alle anderen Eingaben zu “Felder neu skalieren” deaktiviert und die Auswahl beibehalten.

Analysegewichtung. Diese Variable enthält Analysegewichtungen (Regression oder Stichprobe). Analysegewichtungen werden verwendet, um Differenzen in der Varianz zwischen den Ebenen des Zielfelds zu berücksichtigen. Ein stetiges Feld auswählen.

Stetige Eingabefelder. Mit dieser Option werden stetige Eingabefelder durch eine z-Wert-Transformation oder eine Min./Max. Transformation normalisiert. Die Neuskalierung von Eingaben ist besonders nützlich, wenn Sie Funktionserstellung durchführen in den Einstellungen “Auswählen und erstellen” auswählen.

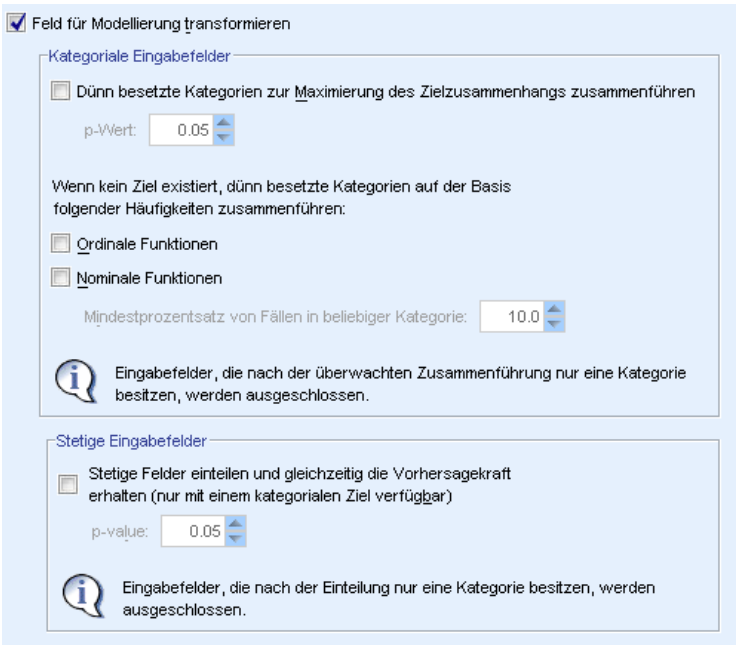
- **Z-Wert-Transformation.** Die Felder werden mithilfe des beobachteten Mittelwerts und der Standardabweichung als Schätzungen der Populationsparameter standardisiert und die z-Werte werden anschließend den entsprechenden Werten einer Normalverteilung mit den Angaben fürEndgültiger Mittelwert und Endgültige Standardabweichung zugeordnet. Geben Sie eine Zahl für Endgültiger Mittelwert und eine positive Zahl für Endgültige Standardabweichung an. Die Standardwerte sind entsprechend der standardisierten Neuskalierung 0 bzw. 1.
- **Min./Max. Transformation.** Die Felder werden mithilfe der beobachteten Mindest- und Höchstwerte als Schätzungen der Populationsparameter den entsprechenden Werten einer Gleichverteilung mit den Angaben für Minimum und Maximum zugeordnet. Geben Sie für Maximum eine Zahl größer als Minimum an.

Stetiges Ziel. Mit dieser Option wird ein stetiges Feld mithilfe der Box-Cox-Transformation in ein Feld transformiert, das eine ungefähre Normalverteilung mit den Angaben für Endgültiger Mittelwert und Endgültige Standardabweichung aufweist. Geben Sie eine Zahl für Endgültiger Mittelwert und eine positive Zahl für Endgültige Standardabweichung an. Die Standardwerte sind 0 bzw. 1.

Hinweis: Wenn ein Ziel durch ADP transformiert wurde, bewerten nachfolgend mithilfe des transformierten Ziels erstellte Modelle die transformierten Einheiten. Um die Ergebnisse interpretieren und verwenden zu können, müssen Sie den vorhergesagten Wert wieder in das ursprüngliche metrische Maß zurückkonvertieren. Für weitere Informationen siehe Transformiert Werte zurück auf S. 43.

Felder transformieren

Abbildung 4-8
Automatisierte Datenaufbereitung – Felder transformieren – Einstellungen


Feld für Modellierung transformieren
Kategoriale Eingabefelder
Dünn besetzte Kategorien zur Maximierung des Zielzusammenhangs zusammenführen
p-Wert: 0.05
Wenn kein Ziel existiert, dünn besetzte Kategorien auf der Basis folgender Häufigkeiten zusammenführen:
Ordinale Funktionen
Nominale Funktionen
Mindestprozentsatz von Fällen in beliebiger Kategorie: 10.0
Eingabefelder, die nach der überwachten Zusammenführung nur eine Kategorie besitzen, werden ausgeschlossen.
Stetige Eingabefelder
Stetige Felder einteilen und gleichzeitig die Vorhersagekraft erhalten (nur mit einem kategorialen Ziel verfügbar)
p-value: 0.05
Eingabefelder, die nach der Einteilung nur eine Kategorie besitzen, werden ausgeschlossen.

Um die Vorhersagekraft Ihrer Daten zu verbessern, können Sie die Eingabefelder transformieren.

Feld für Modellierung transformieren. Durch Deaktivieren dieser Option werden alle anderen Eingaben zu “Felder transformieren” deaktiviert und die Auswahl beibehalten.

Kategoriale Eingabefelder

- **Dünn besetzte Kategorien zur Maximierung des Zielzusammenhangs zusammenführen.** Mit dieser Option erstellen Sie ein sparsameres Modell, indem die Anzahl der zu verarbeitenden Felder in Zusammenhang mit dem Ziel reduziert wird. Ähnliche Kategorien werden anhand der Beziehung zwischen der Eingabe und dem Ziel identifiziert. Kategorien, die sich nicht signifikant unterscheiden (d. h. einen *p*-Wert aufweisen, der größer als der angegebene Wert ist), werden zusammengeführt. Geben Sie einen Wert größer als 0 und kleiner oder gleich 1 an. Wenn alle Kategorien zu einer zusammengeführt werden, werden die Original- und abgeleiteten Versionen des Felds aus der weiteren Analyse ausgeschlossen, da sie keinen Wert als Einflussvariable aufweisen.
- **Wenn kein Ziel existiert, dünn besetzte Kategorien auf der Basis folgender Häufigkeiten zusammenführen.** Wenn das Daten-Set kein Ziel aufweist, können Sie dünn besetzte Kategorien von ordinalen und nominalen Feldern zusammenführen. Die Methode der

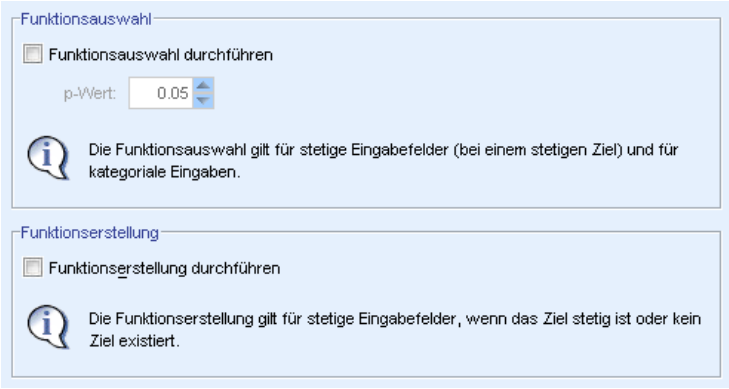
gleichen Häufigkeiten wird verwendet, um Kategorien mit weniger als dem angegebenen Mindestprozentsatz der Gesamtanzahl an Datensätzen zusammenzuführen. Geben Sie einen Wert größer oder gleich 0 und kleiner als 100 ein. Der Standardwert ist 10. Die Zusammenführung wird beendet, wenn keine Kategorien mit weniger als dem angegebenen Mindestprozentsatz an Fällen vorhanden sind oder wenn nur noch zwei Kategorien übrig sind.

Stetige Eingabefelder. Wenn das Daten-Set ein kategoriales Ziel enthält, können Sie stetige Eingaben mit starkem Zusammenhang einteilen, um die Verarbeitungsleistung zu verbessern. Klassen werden anhand der Eigenschaften "homogener Untergruppen" erstellt, die durch die Scheffé-Methode mithilfe des angegebenen *p*-Werts als Alpha für den kritischen Wert zur Bestimmung homogener Untergruppen identifiziert werden. Geben Sie einen Wert größer als 0 und kleiner oder gleich 1 ein. Der Standardwert ist 0,05. Wenn in dem Klassierungsvorgang eine einzelne Klassierung für ein bestimmtes Feld durchgeführt wird, werden die Original- und eingeteilten Versionen des Felds ausgeschlossen, da sie keinen Wert als Einflussvariable aufweisen.

Hinweis: Die Klassierung in ADP unterscheidet sich von der optimalen Klassierung. Bei der optimalen Klassierung werden Entropieinformationen verwendet, um ein stetiges Feld in ein kategoriales Feld umzuwandeln; dazu müssen Daten sortiert und im Arbeitsspeicher abgelegt werden. ADP verwendet homogene Untergruppen zum Klassieren eines stetigen Felds, das bedeutet, dass die ADP-Klassierung keine Daten sortieren und im Arbeitsspeicher ablegen muss. Der Einsatz homogener Untergruppen zum Klassieren eines stetigen Felds bedeutet, dass die Anzahl der Kategorien nach der Klassierung immer kleiner oder gleich der Anzahl der Kategorien im Ziel ist.

Auswählen und erstellen

Abbildung 4-9
Automatisierte Datenaufbereitung – Auswählen und erstellen – Einstellungen


Funktionsauswahl
Funktionsauswahl durchführen
p-Wert:
0.05
Die Funktionsauswahl gilt für stetige Eingabefelder (bei einem stetigen Ziel) und für kategoriale Eingaben.
Funktionserstellung
Funktionserstellung durchführen
Die Funktionserstellung gilt für stetige Eingabefelder, wenn das Ziel stetig ist oder kein Ziel existiert.

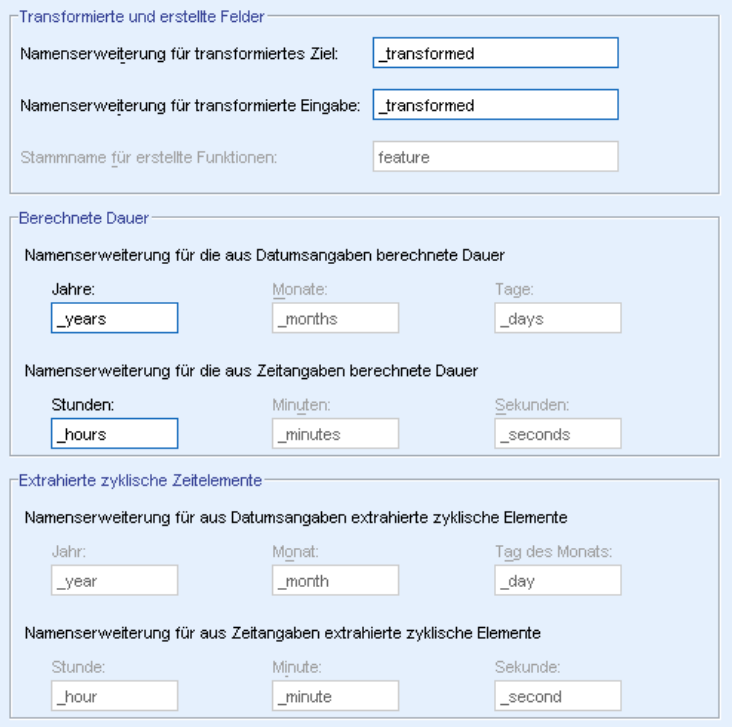
Um die Vorhersagekraft Ihrer Daten zu verbessern, können Sie basierend auf den bestehenden Feldern neue Felder erstellen.

Funktionsauswahl durchführen. Eine stetige Eingabe wird aus der Analyse entfernt, wenn der *p*-Wert für seine Korrelation mit dem Ziel größer ist als der angegebene *p*-Wert.

Funktionserstellung durchführen. Wählen Sie diese Option aus, um neue Funktionen von einer Kombination aus mehreren bestehenden Funktionen abzuleiten. Die alten Funktionen werden bei der weiteren Analyse nicht verwendet. Diese Option gilt nur für stetige Eingabefunktionen mit stetigem Ziel oder Eingabefunktionen, in denen kein Ziel vorhanden ist.

Namensfelder

Abbildung 4-10
Automatisierte Datenaufbereitung – Namensfelder – Einstellungen



Zur einfachen Identifikation neuer und transformierter Funktionen erstellt ADP allgemeine neue Namen, Präfixe oder Suffixe und wendet diese an. Sie können diese Namen ändern und ihnen mehr Aussagekraft für Ihre eigenen Anforderungen und Daten geben.

Transformierte und erstellte Felder. Geben Sie die Namenserweiterungen an, die auf transformierte Ziel- und Eingabefelder angewendet werden sollen.

Geben Sie außerdem über die Einstellungen “Auswählen und erstellen” den Präfixnamen an, der auf erstellte Funktionen angewendet werden soll. Der neue Name wird erstellt, indem ein numerisches Suffix an diesen Präfix-Stammnamen angehängt wird. Das Zahlenformat hängt davon ab, wie viele neue Funktionen abgeleitet werden, zum Beispiel:

- Es werden 1-9 erstellte Funktionen benannt: Funktion1 bis Funktion9.
- Es werden 10-99 erstellte Funktionen benannt: Funktion01 bis Funktion99.
- Es werden 100-999 erstellte Funktionen benannt: Funktion001 bis Funktion999 usw.

So wird gewährleistet, dass die erstellten Funktionen ungeachtet ihrer Anzahl in einer vernünftigen Reihenfolge sortiert werden.

Berechnete Dauer. Geben Sie die Namenserweiterungen an, die auf die aus Datums- und Zeitangaben berechnete Dauer angewendet werden sollen.

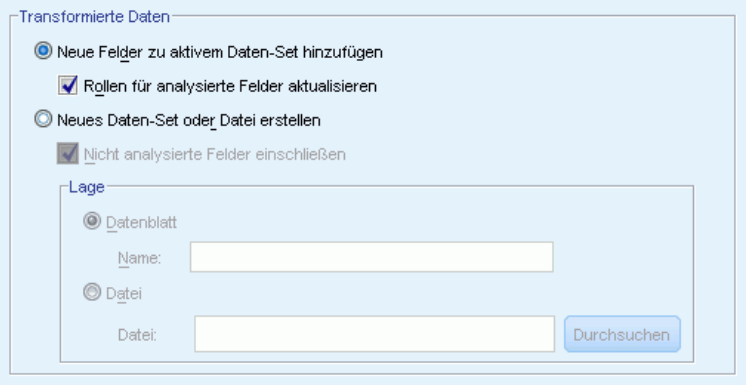
Extrahierte zyklische Zeitelemente. Geben Sie die Namenserweiterungen an, die auf die aus Datums- und Zeitangaben extrahierten zyklischen Elemente angewendet werden sollen.

Transformationen anwenden und speichern

Jenachdem, ob Sie die Dialogfelder für interaktive oder automatische Datenaufbereitung verwenden, weichen die Einstellungen zum Anwenden und Speichern von Transformationen leicht voneinander ab.

Interaktive Datenaufbereitung – Transformationen anwenden – Einstellungen

Abbildung 4-11
Interaktive Datenaufbereitung – Transformationen anwenden – Einstellungen

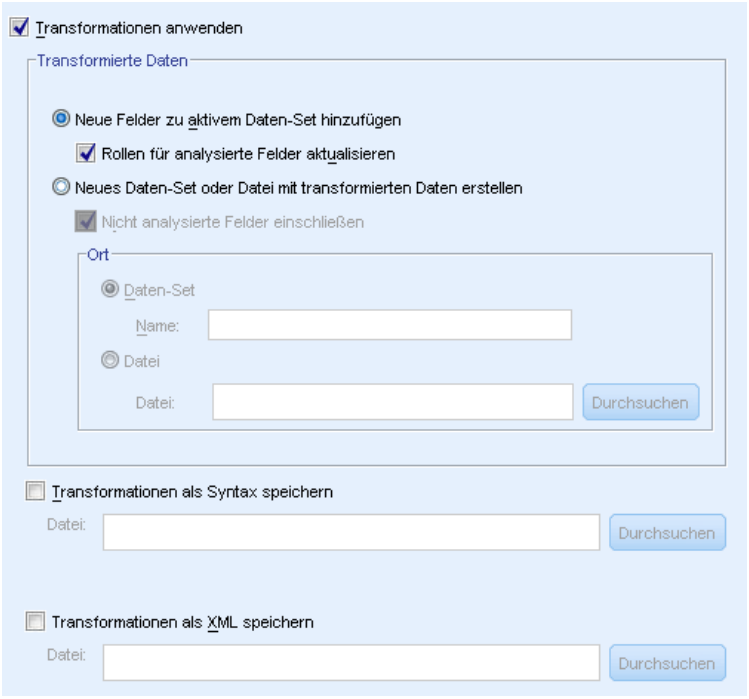


Transformierte Daten. Diese Einstellungen legen den Speicherort der transformierten Daten fest.

- **Neue Felder zu aktivem Daten-Set hinzufügen.** Alle durch die automatisierte Datenaufbereitung erstellten Felder werden dem aktiven Daten-Set als neue Felder hinzugefügt. Mit der Option Rollen für analysierte Felder aktualisieren wird die Rolle für alle Felder, die von der weiteren Analyse durch die automatisierte Datenaufbereitung ausgeschlossen werden, auf "Keine" gesetzt.
- **Neues Daten-Set oder Datei mit transformierten Daten erstellen.** Von der automatisierten Datenaufbereitung empfohlene Felder werden einem neuen Daten-Set oder einer Datei hinzugefügt. Mit der Option Nicht analysierte Felder einschließen werden dem Original-Daten-Set Felder hinzugefügt, die im neuen Daten-Set auf der Registerkarte "Felder" nicht angegeben wurden. Das ist nützlich beim Übertragen von Feldern, die Informationen enthalten, die bei der Modellierung nicht verwendet werden, wie zum Beispiel eine ID, eine Adresse oder ein Name, in das neue Daten-Set.

Automatische Datenaufbereitung – Anwenden und speichern – Einstellungen

Abbildung 4-12
Automatische Datenaufbereitung – Anwenden und speichern – Einstellungen



Die Gruppe “Transformierte Daten” ist dieselbe wie in der interaktiven Datenaufbereitung. Bei der automatischen Datenaufbereitung sind die folgenden zusätzlichen Optionen verfügbar:

Transformationen anwenden. Wird im Dialogfeld der automatischen Datenaufbereitung diese Option deaktiviert, werden alle anderen Befehle “Anwenden und speichern” deaktviert und die Auswahl beibehalten.

Transformationen als Syntax speichern. Mit dieser Option werden die empfohlenen Transformationen als Befehlssyntax in eine externe Datei gespeichert. Das Dialogfeld “Interaktive Datenaufbereitung” enthält diese Steuerung nicht, da es die Transformationen als Befehlssyntax in das Syntaxfenster einfügt, wenn Sie auf Einfügen klicken.

Transformationen als XML speichern. Mit dieser Option werden die empfohlenen Transformationen als XML in einer externen Datei gespeichert, die mithilfe von `TMS MERGE` mit der Modell-PMML zusammengeführt oder mithilfe von `TMS IMPORT` auf ein anderes Daten-Set angewendet werden kann. Das Dialogfeld “Interaktive Datenaufbereitung” enthält diese Steuerung nicht, da es die Transformationen als XML speichert, wenn Sie in der Symbolleiste im oberen Bereich des Dialogfelds auf XML speichern klicken.

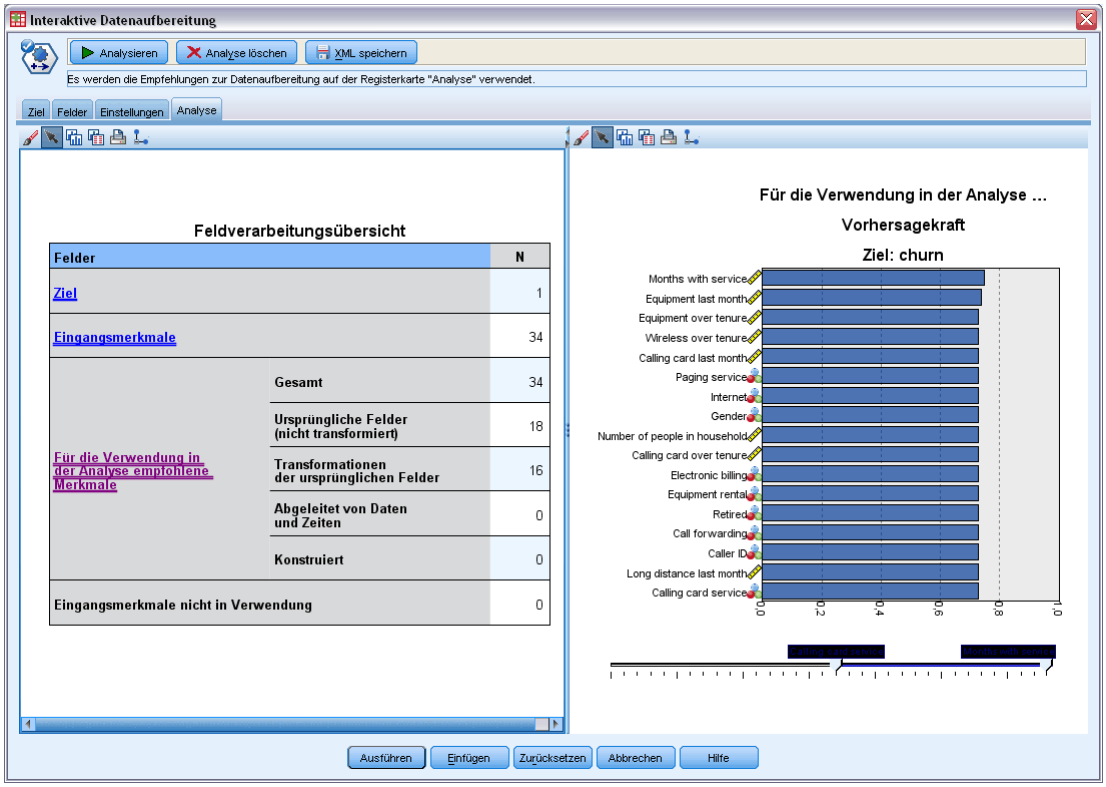
Registerkarte "Analyse"

Anmerkung: Die Registerkarte "Analyse" wird in der interaktiven Datenaufbereitung verwendet, damit Sie die empfohlenen Transformationen überprüfen können. Das Dialogfeld "Automatische Datenaufbereitung" enthält diesen Schritt nicht.

- Wenn Sie mit den ADP-Einstellungen einschließlich aller in den Registerkarten "Ziel", "Felder" und "Einstellungen" vorgenommenen Änderungen zufrieden sind, klicken Sie auf Daten analysieren. Der Algorithmus wendet die Eingabedaten an und zeigt die Ergebnisse auf der Registerkarte "Analyse" an.

Die Registerkarte "Analyse" enthält Ausgaben in Grafik- und Tabellenform, die die Verarbeitung Ihrer Daten zusammenfassen, und zeigt Empfehlungen an, wie die Daten möglicherweise bearbeitet oder zum Scoring verbessert werden können. Anschließend können Sie diese Empfehlungen überprüfen und entweder akzeptieren oder ablehnen.

Abbildung 4-13
Registerkarte "Analyse" in der automatisierten Datenaufbereitung



Die Registerkarte "Analyse" besteht aus zwei Bereichen, der Hauptansicht im linken Bereich und der verknüpften oder Hilfsansicht im rechten Bereich. Es gibt drei Hauptansichten:

- Feldverarbeitungsübersicht (Standard). Für weitere Informationen siehe Feldverarbeitungsübersicht auf S. 32.
- Felder. Für weitere Informationen siehe Felder auf S. 33.
- Aktionsübersicht. Für weitere Informationen siehe Aktionsübersicht auf S. 35.

Es gibt vier verknüpfte/Hilfsansichten:

- Vorhersagekraft (Standard). Für weitere Informationen siehe Vorhersagekraft auf S. 36.
- Feldertabelle. Für weitere Informationen siehe Feldertabelle auf S. 37.
- Felddetails. Für weitere Informationen siehe Felddetails auf S. 38.
- Aktionsdetails. Für weitere Informationen siehe Aktionsdetails auf S. 40.

Verknüpfungen zwischen Ansichten

In der Hauptansicht steuert unterstrichener Text in den Tabellen die Anzeige in der verknüpften Ansicht. Wenn Sie auf den Text klicken, erhalten Sie Informationen über ein bestimmtes Feld, ein Set von Feldern oder einen Verarbeitungsschritt. Der zuletzt von Ihnen ausgewählte Link wird in einer dunkleren Farbe angezeigt; dies hilft Ihnen dabei, die Verbindung zwischen den Inhalten der beiden Ansichtsbereiche zu identifizieren.

Zurücksetzen der Ansichten

Klicken Sie auf Zurücksetzen im unteren Bereich der Hauptansicht, um die ursprünglichen Empfehlungen der Analyse erneut anzuzeigen und alle in den Analyseansichten vorgenommenen Änderungen rückgängig zu machen.

Feldverarbeitungsübersicht

Abbildung 4-14
Feldverarbeitungsübersicht

Felder		N
Ziel		1
Eingangsmerkmale		6
Für die Verwendung in der Analyse empfohlene Merkmale	Gesamt	2
	Ursprüngliche Felder (nicht transformiert)	0
	Transformationen der ursprünglichen Felder	2
	Abgeleitet von Daten und Zeiten	0
	Konstruiert	0
Eingangsmerkmale nicht in Verwendung		4

Die Tabelle “Feldverarbeitungsübersicht” gibt Ihnen eine Momentaufnahme des projizierten Gesamteinflusses der Verarbeitung, einschließlich Änderungen des Status der Funktionen und der Anzahl der erstellten Funktionen.

Beachten Sie, dass dabei kein Modell erstellt wird und somit kein Maß oder keine Grafik der Veränderung der Gesamtvorhersagekraft vor und nach der Datenaufbereitung vorhanden ist; Sie können stattdessen Grafiken der Vorhersagekraft einzelner empfohlener Einflussvariablen anzeigen.

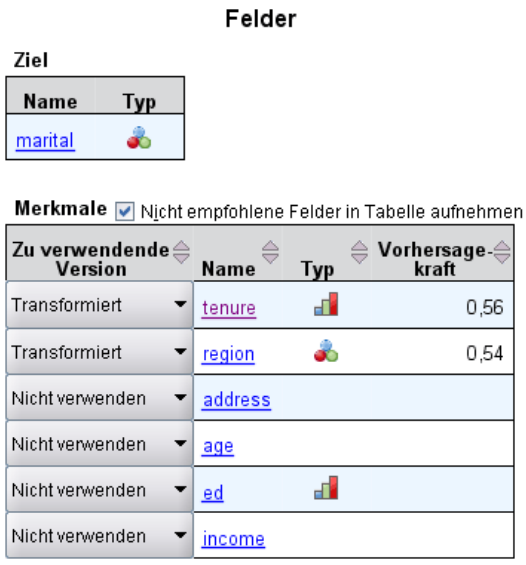
Die Tabelle zeigt folgende Informationen an:

- Die Anzahl der Zielfelder.
- Die Anzahl der ursprünglichen Eingabefelder.
- Die für die Analyse und die Modellierung empfohlenen Felder. Dazu zählen die Gesamtanzahl der empfohlenen Felder, die Anzahl der ursprünglichen untransformierten Felder, die Anzahl der transformierten Felder (ausgenommen Zwischenversionen von aus Datums-/Zeitfeldern abgeleiteten Feldern und erstellte Felder), die Anzahl der aus Datums-/Zeitfeldern abgeleiteten Felder und die Anzahl der erstellten Felder.
- Die Anzahl der Eingabefelder, die in keiner Form empfohlen werden, sei es in ihrer ursprünglichen Form, als abgeleitetes Feld oder als Eingabe in einer erstellten Funktion.

Klicken Sie auf die unterstrichenen Informationen unter Felder, um weitere Informationen in einer verknüpften Ansicht anzuzeigen. In der verknüpften Ansicht "Feldertabelle" erhalten Sie Informationen über Ziel, Eingabefunktionen und Nicht verwendte Eingabefunktionen. Für weitere Informationen siehe Feldertabelle auf S. 37. Empfohlene Funktionen für den Einsatz in Analysen werden in der verknüpften Ansicht "Vorhersagekraft" angezeigt. Für weitere Informationen siehe Vorhersagekraft auf S. 36.

Felder

Abbildung 4-15
Felder



In der Hauptansicht "Felder" werden die verarbeiteten Felder angezeigt sowie, ob ADP diese zur Verwendung in nachgelagerten Modellen empfiehlt. Sie können die Empfehlung für jedes Feld überschreiben, zum Beispiel, um erstellte Funktionen auszuschließen oder Funktionen einzuschließen, von denen ADP empfiehlt, sie auszuschließen. Wenn ein Feld transformiert wurde, können Sie entscheiden, ob Sie die vorgeschlagene Transformation akzeptieren oder die Originalversion verwenden möchten.

Die Felderansicht besteht aus zwei Tabellen, eine für das Ziel und eine für Funktionen, die entweder verarbeitet oder erstellt wurden.

Tabelle "Ziel"

Die Tabelle Ziel wird nur angezeigt, wenn in den Daten ein Ziel definiert wurde.

Die Tabelle enthält zwei Spalten:

- **Name.** Dies ist der Name oder die Bezeichnung des Zielfelds. Der Originalname wird immer verwendet, auch wenn das Feld transformiert wurde.
- **Typ.** Hier erscheint das Symbol für das entsprechende Messniveau; fahren Sie mit der Maus über das Symbol, um eine Bezeichnung (stetig, sortiertes Set, Set usw.) anzuzeigen, die die Daten beschreibt.

  Wenn das Ziel transformiert wurde, gibt die Spalte Typ die endgültige transformierte Version an. *Hinweis*: Transformationen für das Ziel können nicht abgeschaltet werden.

Tabelle "Funktionen"

Die Tabelle Funktionen wird immer angezeigt. Jede Zeile der Tabelle repräsentiert ein Feld. Standardmäßig sind die Zeilen nach absteigender Vorhersagekraft sortiert.

Bei gewöhnlichen Funktionen wird der Originalname immer als Zeilenname verwendet. Sowohl Original- als auch abgeleitete Versionen von Datums-/Zeitfeldern werden in der Tabelle (in getrennten Zeilen) angezeigt; die Tabelle enthält auch erstellte Funktionen.

Beachten Sie, dass transformierte Versionen von in der Tabelle angezeigten Feldern immer die Endversionen darstellen.

Die Tabelle enthält folgende Spalten:

- **Zu verwendende Version.** Hier wird eine Dropdown-Liste angezeigt, die festlegt, ob ein Feld nachgelagert verwendet wird oder ob die vorgeschlagenen Transformationen verwendet werden sollen. Standardmäßig werden in der Dropdown-Liste die Empfehlungen wiedergegeben.

  Für gewöhnliche Funktionen, die transformiert wurden, stehen in der Dropdown-Liste drei Optionen zur Auswahl: Transformiert, Original und Nicht verwenden.

  Für nicht transformierte gewöhnliche Funktionen sind folgende Auswahlmöglichkeiten verfügbar: Original und Nicht verwenden.

  Für abgeleitete Datums-/Zeitfelder und erstellte Funktionen sind folgende Auswahlmöglichkeiten verfügbar: Transformiert und Nicht verwenden.

  Für Original-Datumsfelder ist die Dropdown-Liste deaktiviert und auf Nicht verwenden gesetzt.

  Hinweis: Für Funktionen mit Original- und transformierten Versionen werden bei einem Wechsel zwischen den Versionen Original und Transformiert automatisch die Einstellungen Typ und Vorhersagekraft für diese Funktionen aktualisiert.

- **Name.** Jeder Feldname ist ein Link. Klicken Sie auf den Namen, um in der verknüpften Ansicht weitere Informationen über das Feld anzuzeigen. Für weitere Informationen siehe Felddetails auf S. 38.
- **Typ.** Hier erscheint das Symbol für den entsprechenden Datentyp; fahren Sie mit der Maus über das Symbol, um eine Bezeichnung (stetig, sortiertes Set, Set usw.) anzuzeigen, die die Daten beschreibt.
- **Vorhersagekraft.** Die Vorhersagekraft wird nur für Felder angezeigt, die von ADP empfohlen werden. Diese Spalte wird nicht angezeigt, wenn kein Ziel definiert wurde. Die Vorhersagekraft reicht von 0 bis 1, wobei größere Werte "bessere" Einflussgrößen andeuten. Im Allgemeinen ist die Vorhersagekraft für den Vergleich von Einflussgrößen in einer ADP-Analyse nützlich, doch sollten Vorhersagekraft-Werte nicht in Analysen verglichen werden.

Standardmäßig werden in der Tabelle "Funktionen" nur empfohlene Felder angezeigt. Um die restlichen Felder anzuzeigen, wählen Sie das Feld Nicht empfohlene Felder in Tabelle einschließen über der Tabelle aus; diese Felder werden dann am Ende der Tabelle angezeigt.

Aktionsübersicht

Abbildung 4-16
Aktionsübersicht

Bei jeder von der automatisierten Datenaufbereitung vorgenommenen Aktion werden Eingabefunktionen tranformiert und/oder herausgefiltert. Felder, die in einer Aktion erhalten bleiben, werden in der nächsten verwendet. Die Felder, die bis zum letzten Schritt erhalten

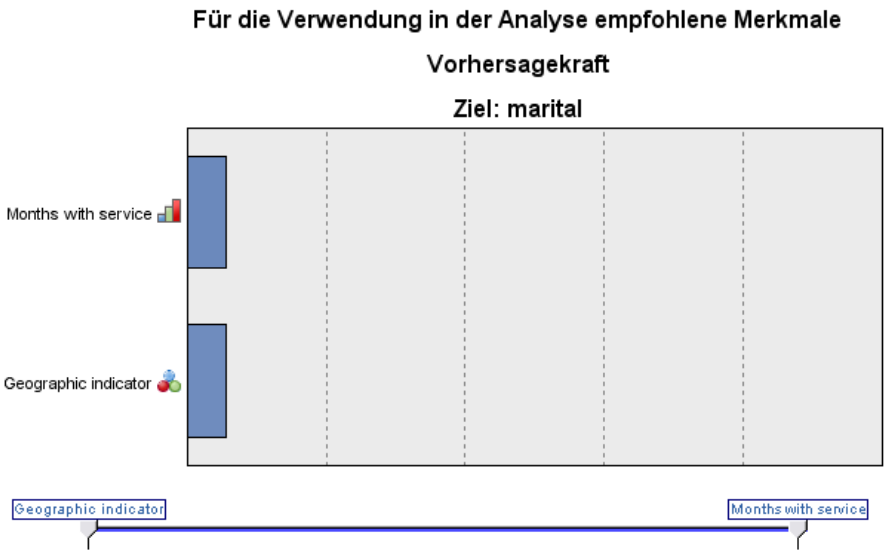
bleiben, werden dann für die Modellierung empfohlen, während Eingaben zu transformierten und erstellten Funktionen durch Filterung ausgeschlossen werden.

Die Aktionsübersicht ist eine einfache Tabelle, in der die von der ADP vorgenommenen Verarbeitungsaktionen aufgelistet sind. Klicken Sie auf den unterstrichenen Link Beschreibung, um in einer verknüpften Ansicht weitere Informationen über die durchgeführten Schritte anzuzeigen. Für weitere Informationen siehe Aktionsdetails auf S. 40.

Hinweis: Es werden nur die Original- und endgültigen transformierten Versionen jedes Felds angezeigt, jedoch keine während der Analyse verwendeten Zwischenversionen.

Vorhersagekraft

Abbildung 4-17
Vorhersagekraft



Wird standardmäßig bei der ersten Ausführung der Analyse angezeigt. Wenn Sie dagegen Empfohlene Funktionen für den Einsatz in Analysen in der Hauptansicht “Feldverarbeitungsübersicht” auswählen, zeigt das Diagramm die Vorhersagekraft der empfohlenen Funktionen an. Felder werden nach Vorhersagekraft sortiert, wobei das Feld mit dem höchsten Wert zuerst erscheint.

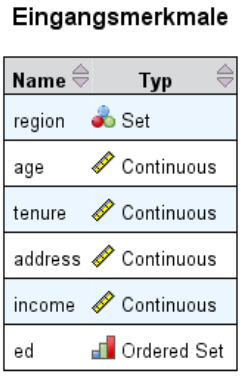
Bei transformierten Versionen gewöhnlicher Funktionen, gibt der Feldname Ihre Suffixauswahl im Bereich “Feldnamen” auf der Registerkarte “Einstellungen” an, zum Beispiel: *_transformiert*.

Diese Symbole werden nach den einzelnen Feldnamen angezeigt.

Die Vorhersagekraft jeder empfohlenen Funktion oder Einflussvariablen wird entweder aus einer linearen Regression oder einem Naive Bayes-Modell berechnet, abhängig davon, ob das Ziel stetig oder kategorial ist.

Feldertabelle

Abbildung 4-18
Feldertabelle



Die Feldertabelle wird angezeigt, wenn Sie in der Hauptansicht "Feldverarbeitungsübersicht" auf Ziel, Eingabefunktionen oder Nicht verwendete Eingabefunktionen klicken, und enthält eine einfache Tabelle, die die wichtigsten Funktionen auflistet.

Die Tabelle enthält zwei Spalten:

- **Name.** Der Funktions- oder Feldname.

  Für Ziele wird der Originalname oder die Originalbeschriftung verwendet, selbst wenn das Ziel transformiert wurde.

  Bei transformierten Versionen gewöhnlicher Funktionen gibt der Name Ihre Suffixauswahl im Bereich "Feldnamen" auf der Registerkarte "Einstellungen" an, zum Beispiel: *_transformiert*.

  Bei aus Datums- und Zeitangaben abgeleiteten Feldern wird der Name der endgültigen transformierten Version verwendet, zum Beispiel: *bdatum_Jahre*.

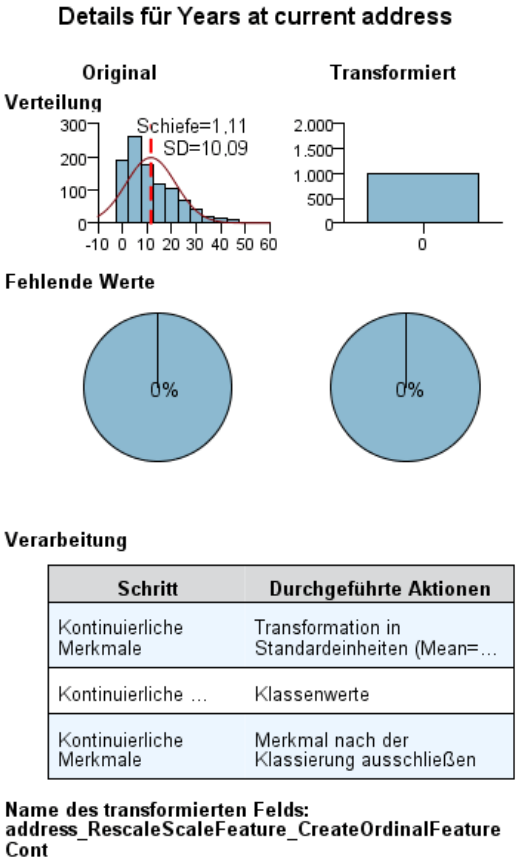
  Bei erstellten Funktionen wird der Name der erstellten Funktion verwendet, zum Beispiel: *Funktion1*.

- **Typ.** Hier erscheint das Symbol für den entsprechenden Datentyp.

  Für das Ziel gibt der Typ stets die transformierte Version wieder (wenn das Ziel transformiert wurde), zum Beispiel bei einem Wechsel von ordinal (sortiertes Set) zu stetig (Bereich, Skala) oder umgekehrt.

Felddetails

Abbildung 4-19
Felddetails



Die Ansicht “Felddetails” wird angezeigt, wenn Sie auf Name in der Hauptansicht “Felder” klicken, und enthält Informationen über Verteilung, fehlende Werte und (falls zutreffend) Vorhersagekraft-Diagramme für das ausgewählte Feld. Außerdem wird der Verarbeitungsverlauf für das Feld und der Name des transformierten Felds angezeigt (falls zutreffend).

Für jedes Diagramm-Set werden nebeneinander zwei Versionen angezeigt, um das Feld mit und ohne angewendete Transformationen zu vergleichen. Wenn keine transformierte Version des Felds vorhanden ist, wird nur ein Diagramm für die Originalversion angezeigt. Für abgeleitete Datums- und Zeitfelder sowie erstellte Funktionen werden die Diagramme nur für die neue Funktion angezeigt.

Hinweis: Wenn ein Feld wegen zu vieler Kategorien ausgeschlossen wurde, wird nur der Verarbeitungsverlauf angezeigt.

Verteilungsdiagramm

Die Verteilung stetiger Felder wird als Histogramm angezeigt, mit einer überlagerten Normalverteilungskurve und einer vertikalen Referenzlinie für den Mittelwert; kategoriale Felder werden als Balkendiagramm angezeigt.

Die Histogramme werden nach Standardabweichung und Schiefe bezeichnet, allerdings wird Letztere nicht angezeigt, wenn die Anzahl der Werte kleiner gleich 2 oder die Varianz des originalen Felds kleiner als 10-20 ist.

Fahren Sie mit der Maus über das Diagramm, um entweder den Mittelwert für Histogramme oder die Zählung und den Prozentsatz der Gesamtzahl der Datensätze für Kategorien in Balkendiagrammen anzuzeigen.

Diagramm fehlender Werte

Kreisdiagramme vergleichen den Prozentsatz fehlender Werte mit und ohne angewendete Transformationen; die Diagrammbeschriftungen zeigen den Prozentsatz an.

Wenn ADP die Behandlung fehlender Werte durchgeführt hat, enthält das Kreisdiagramm nach der Transformation auch den Ersatzwert als Beschriftung, d. h. den anstelle von fehlenden Werten verwendeten Wert.

Fahren Sie mit der Maus über das Diagramm, um die Zählung der fehlenden Werte und den Prozentsatz der Gesamtzahl an Datensätzen anzuzeigen.

Vorhersagekraft-Diagramme

Für empfohlene Felder zeigen Balkendiagramme die Vorhersagekraft vor und nach der Transformation an. Wenn das Ziel transformiert wurde, steht die berechnete Vorhersagekraft in Beziehung zum transformierten Ziel.

Hinweis: Die Vorhersagekraft-Diagramme werden nicht angezeigt, wenn kein Ziel definiert wurde oder wenn Sie in der Hauptansicht auf das Ziel klicken.

Fahren Sie mit der Maus über das Diagramm, um den Wert der Vorhersagekraft anzuzeigen.

Tabelle "Verarbeitungsverlauf"

Die Tabelle zeigt, wie die transformierte Version eines Felds abgeleitet wurde. Von ADP durchgeführte Aktionen werden in der Reihenfolge ihrer Ausführung aufgelistet. Bei bestimmten Schritten wurden jedoch unter Umständen mehrere Aktionen für ein spezielles Feld durchgeführt.

Hinweis: Die Tabelle wird nur für transformierte Felder angezeigt.

Die Informationen in der Tabelle erscheinen in drei Spalten:

- **Stufen.** Der Name der Stufe. Zum Beispiel "Stetige Funktionen". Für weitere Informationen siehe Aktionsdetails auf S. 40.

- **Durchgeführte Aktionen.** Die Liste der durchgeführten Verarbeitung. Zum Beispiel “Zu Standardeinheiten transformieren”.
- **Funktion.** Diese Spalte erscheint nur bei erstellten Funktionen und zeigt die lineare Kombination von Eingabefeldern an, zum Beispiel 0,06*Alter + 1,21*Größe.

Aktionsdetails

Abbildung 4-20
ADP-Analyse – Aktionsdetails

Stetige Funktionen

Transformation	Anzahl der Funktionen	Kriterien	
		Mittelwert	SD
In Standardeinheiten transformieren	16	0,00	1,00

Erstellung eines Funktionsbereichs	X
Erstellte Funktionen	0
Funktionen, die wegen niedrigem Zielzusammenhang ausgeschlossen wurden	0
Funktionen, die ausgeschlossen wurden, weil sie nach der Einteilung konstant waren.	0

Die verknüpfte Ansicht “Aktionsdetails” wird angezeigt, wenn Sie in der Hauptansicht “Aktionsübersicht” auf den unterstrichenen Link Beschreibung klicken, und enthält sowohl aktionsspezifische als auch allgemeine Informationen über jeden durchgeführten Verarbeitungsschritt. Die aktionsspezifischen Informationen erscheinen stets zuerst.

Für jede Aktion wird die Beschreibung als Titel im oberen Bereich der verknüpften Ansicht verwendet. Die aktionsspezifischen Informationen erscheinen unter dem Titel und enthalten unter Umständen Details zur Anzahl abgeleiteter Funktionen, zu umgewandelten Feldern, zu Zieltransformationen, zu zusammengeführten oder neu sortierten Kategorien und zu erstellten oder ausgeschlossenen Funktionen.

Bei der Verarbeitung jeder Aktion kann sich die für die Verarbeitung verwendete Anzahl an Funktionen ändern, wenn beispielsweise Funktionen ausgeschlossen oder zusammengeführt werden.

Hinweis: Wenn eine Aktion deaktiviert oder kein Ziel angegeben wurde, erscheint eine Fehlermeldung anstelle der Aktionsdetails, wenn Sie in der Hauptansicht “Aktionsübersicht” auf den Schritt klicken.

Es gibt neun mögliche Aktionen, davon sind allerdings nicht alle notwendigerweise für jede Analyse aktiv.

Tabelle "Textfelder"

Die Tabelle zeigt folgende Anzahl:

- Von der Analyse ausgeschlossene Funktionen.

Tabelle "Datums- und Uhrzeitfunktionen"

Die Tabelle zeigt folgende Anzahl:

- Aus Datums- und Uhrzeitfunktionen abgeleitete Dauer.
- Datums- und Uhrzeitelemente.
- Insgesamt abgeleitete Datums- und Uhrzeitfunktionen.

Das Referenzdatum oder die -uhrzeit wird als Fußnote angezeigt, wenn eine Datumsdauer berechnet wurde.

Tabelle "Funktions-Screening"

Die Tabelle zeigt die Anzahl folgender von der Verarbeitung ausgeschlossener Funktionen:

- Konstanten.
- Felder mit zu vielen fehlenden Werten.
- Funktionen mit zu vielen Fällen in einer einzelnen Kategorie.
- Nominale Felder (Sets) mit zu vielen Kategorien.
- Insgesamt ausgeschlossene Funktionen.

Tabelle "Typ überprüfen"

Die Tabelle zeigt die Anzahl umgewandelter Felder und teilt sich wie folgt auf:

- In stetige Feldern umgewandelte ordinale Felder (sortierte Sets).
- In ordinale Felder (sortierte Sets) umgewandelte stetige Felder.
- Anzahl an Umwandlungen insgesamt.

Wenn keine Felder (Ziel oder Funktionen) stetig oder ordinal (sortierte Sets) waren, wird dies als Fußnote vermerkt.

Tabelle "Ausreißer"

Die Tabelle zeigt, ob und wie Ausreißer behandelt wurden.

- Entweder die Anzahl stetiger Felder, für die Ausreißer gefunden und entfernt wurden, oder die Anzahl stetiger Felder, für die Ausreißer gefunden und als fehlend eingestuft wurden, je nach Ihren Einstellungen im Feld “Eingaben & Ziel vorbereiten” auf der Registerkarte “Einstellungen”.
- Die Anzahl stetiger Felder, die ausgeschlossen wurden, weil sie nach der Ausreißer-Behandlung konstant waren.

Der Ausreißer-Trennwert wird in einer Fußnote vermerkt. Eine weitere Fußnote wird angezeigt, wenn keine Eingabefelder (Ziel oder Funktionen) stetig waren.

Tabelle “Fehlende Werte”

Die Tabelle zeigt die Anzahl an Feldern, in denen fehlende Werte ersetzt wurden, und teilt sich wie folgt auf:

- Ziel. Diese Zeile wird nicht angezeigt, wenn kein Ziel angegeben wurde.
- Merkmale. Dies teilt sich weiter auf in Anzahl an “nominal (Set)”, “ordinal (sortiertes Set)” und “stetig”.
- Die gesamte Anzahl ersetzter fehlender Werte.

Tabelle “Ziel”

Die Tabelle zeigt wie folgt, ob das Ziel tranformiert wurde:

- Box-Cox-Transformation in Normalverteilung. Dies teilt sich weiter in Spalten auf, die die angegebenen Kriterien (Mittelwert und Standardabweichung) und Lambda zeigen.
- Zielkategorien zur Verbesserung der Stabilität neu sortiert.

Tabelle “Kategoriale Funktionen”

Die Tabelle zeigt folgende Anzahl kategorialer Funktionen:

- Wessen Kategorien wurden zur Verbesserung der Stabilität neu sortiert.
- Wessen Kategorien wurden zur Maximierung des Zielzusammenhangs zusammengeführt.
- Wessen Kategorien wurden zur Behandlung dünn besetzter Kategorien zusammengeführt.
- Wegen niedrigem Zielzusammenhang ausgeschlossen.
- Ausgeschlossen, weil nach der Zusammenführung konstant.

Wenn es keine kategorialen Eingabefelder gab, wird dies durch eine Fußnote vermerkt.

Tabelle “Stetige Funktionen”

Es gibt zwei Tabellen. Die erste zeigt eine der folgenden Transformationen:

- Zu Standardeinheiten transformierte Funktionswerte. Zusätzlich werden hier die Anzahl transformierter Funktionen, der angegebene Mittelwert und die Standardabweichung angezeigt.

- Einem gemeinsamen Bereich zugeordnete Funktionswerte. Zusätzlich werden hier die Anzahl der mithilfe der min./max. Transformation transformierten Funktionen sowie die angegebenen Mindest- und Höchstwerte angezeigt.
- Klassierte Funktionswerte und die Anzahl klassierter Funktionen.

Die zweite Tabelle enthält Informationen über die Funktionserstellung, die als Anzahl folgender Funktionen angezeigt werden:

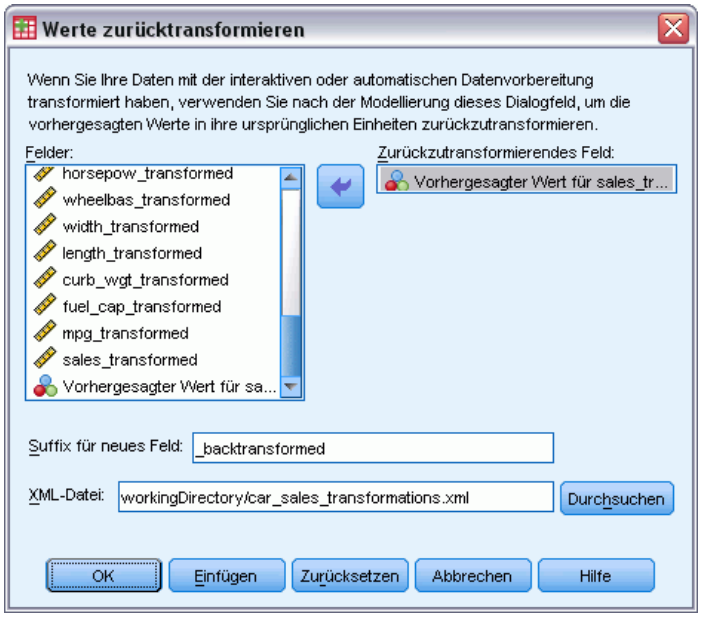
- Erstellt.
- Wegen niedrigem Zielzusammenhang ausgeschlossen.
- Ausgeschlossen, weil nach der Klassierung konstant.
- Ausgeschlossen, weil nach der Erstellung konstant.

Wenn keine stetigen Funktionen eingegeben wurden, wird dies durch eine Fußnote vermerkt.

Transformiert Werte zurück

Wenn ein Ziel durch ADP transformiert wurde, bewerten nachfolgend mithilfe des transformierten Ziels erstellte Modelle die transformierten Einheiten. Um die Ergebnisse interpretieren und verwenden zu können, müssen Sie den vorhergesagten Wert wieder in das ursprüngliche metrische Maß zurückkonvertieren.

Abbildung 4-21
Transformiert Werte zurück



Wählen Sie die folgenden Befehle aus den Menüs aus, um Werte zurückzutransformieren:

Transformieren
Daten für Modellierung vorbereiten
Werte zurücktransformieren...

- Wählen Sie ein Feld, das zurücktransformiert werden soll. Dieses Feld sollte vom Modell vorhergesagte Werte des transformierten Ziels enthalten.
- Geben Sie ein Suffix für das neue Feld an. Dieses neue Feld enthält vom Modell vorhergesagte Werte im ursprünglichen metrischen Maß des nicht transformierten Ziels.
- Geben Sie den Speicherort der XML-Datei mit den ADP-Transformationen an. Es sollte eine Datei sein, die aus den Dialogfeldern für interaktive oder automatische Datenaufbereitung heraus gespeichert wurde. Für weitere Informationen siehe Transformationen anwenden und speichern auf S. 29.

Kapitel
5

Ungewöhnliche Fälle identifizieren

Die Prozedur “Anomalie-Erkennung” sucht anhand von Abweichungen von den Normwerten der Gruppe nach ungewöhnlichen Fällen. Die Prozedur wurde für die Datenprüfung in der explorativen Datenanalyse konzipiert. Zweck der Prozedur ist das schnelle Erkennen von ungewöhnlichen Fällen, bevor mit anderen Analysen Schlüsse aus den Daten gezogen werden. Dieser Algorithmus dient der Erkennung von allgemeinen Anomalien. Dies bedeutet, dass sich die Definition eines anomalen Falls nicht auf eine bestimmte Anwendung beschränkt, bei der Anomalien sehr treffend definiert werden können, z. B. beim Erkennen von ungewöhnlichen Zahlungsmustern im Gesundheitswesen oder beim Aufdecken von Geldwäsche im Finanzwesen.

Beispiel. Ein Analytiker, der mit der Erstellung von Prognosemodellen für die Ergebnisse von Schlaganfallbehandlungen betraut wurde, ist über die Qualität der Daten besorgt, weil solche Modelle bei ungewöhnlichen Beobachtungen anfällig sein können. Einige dieser Randbeobachtungen stellen wirklich einzigartige Fälle dar und eignen sich deswegen nicht für eine Vorhersage. Andere Beobachtungen stellen Dateneingabefehler dar, wobei die Werte technisch gesehen “richtig” sind und deswegen nicht mit Datenvalidierungsprozeduren abgefangen werden können. Die Prozedur “Ungewöhnliche Fälle identifizieren” sucht Ausreißer und meldet diese, sodass der Analytiker entscheiden kann, wie mit diesen Fällen verfahren wird.

Statistiken. Die Prozedur erzeugt Gruppen, Normwerte für Gruppen bei stetigen und kategorialen Variablen, Anomalie-Indizes auf der Grundlage von Abweichungen von den Normwerten der Gruppen sowie Variablen-Einflusswerte für Variablen, die am meisten dazu beitragen, dass ein Falls als ungewöhnlich klassifiziert wird.

Erläuterung der Daten

Daten. Mit dieser Prozedur können sowohl stetige als auch kategoriale Variablen analysiert werden. Jede Zeile stellt eine eindeutige Beobachtung und jede Zeile eine eindeutige Variable als Grundlage für die Gruppen dar. In der Datendatei kann eine Fallidentifizierungsvariable zum Markieren der Ausgabe verfügbar sein. Diese Variable wird jedoch nicht in der Analyse verwendet. Fehlende Werte sind zulässig. Wenn die Gewichtungsvariable angegeben wurde, wird diese ignoriert.

Das Erkennungsmodell kann auf eine neue Test-Datendatei angewendet werden. Die Elemente der Testdaten müssen dieselben wie die Elemente der Lerndaten sein. Abhängig von den Einstellungen des Algorithmus kann die Verarbeitung fehlender Werte, die beim Erstellen des Modells verwendet wird, vor der Bewertung auf die Testdaten angewendet werden.

Fallreihenfolge. Beachten Sie, dass die Lösung von der Fallreihenfolge abhängen kann. Um die Auswirkungen der Reihenfolge zu minimieren, mischen Sie die Fälle in zufälliger Reihenfolge. Prüfen Sie daher die Stabilität einer bestimmten Lösung, indem Sie verschiedene Lösungen

abrufen, bei denen die Fälle in einer unterschiedlichen, zufällig ausgewählten Reihenfolgen sortiert sind. In Situationen mit extrem umfangreichen Dateien können mehrere Durchgänge mit jeweils einer Stichprobe von Fällen durchgeführt werden, die in unterschiedlicher, zufällig ausgewählter Reihenfolge sortiert ist.

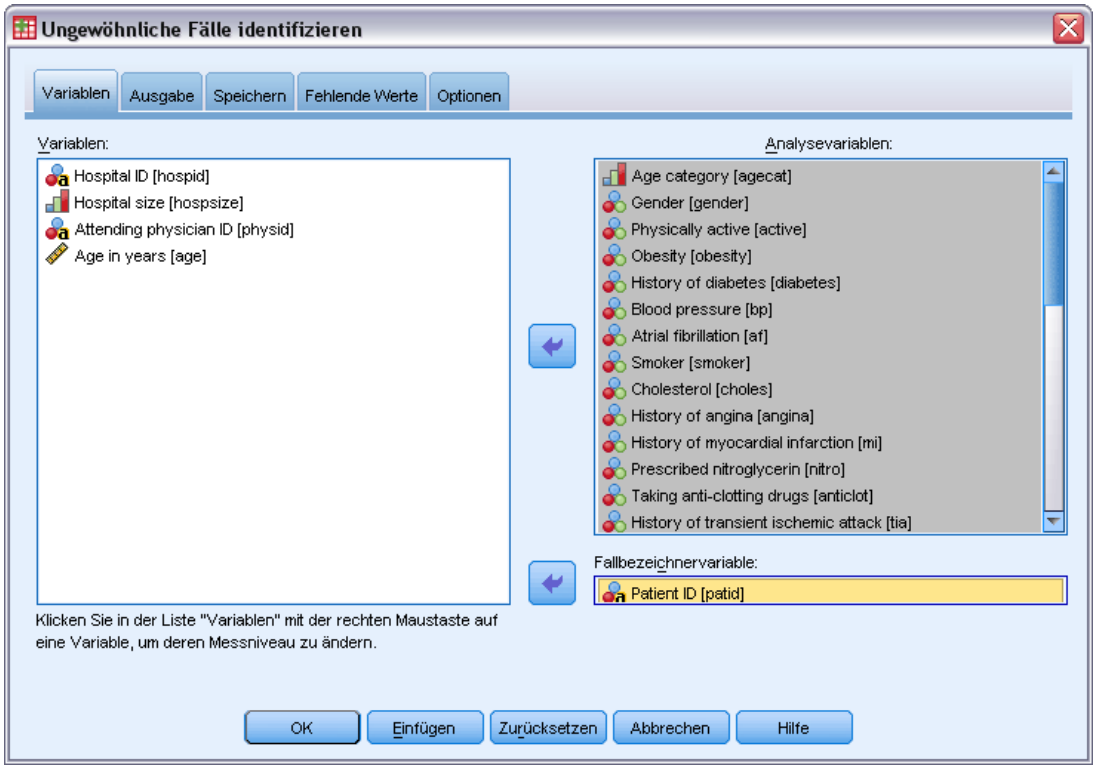
Annahmen. Der Algorithmus setzt voraus, dass alle Variablen nichtkonstant und unabhängig sind. Es wird außerdem angenommen, dass kein Fall bei einer Eingabevariablen fehlende Werte aufweist. Für alle stetigen Variablen wird eine Normalverteilung (Gauß-Verteilung) und für alle kategorialen Variablen eine multinomiale Verteilung vorausgesetzt. Empirische interne Tests zeigen, dass die Prozedur wenig anfällig gegenüber Verletzungen hinsichtlich der Unabhängigkeitsannahme und der Verteilungsannahme ist. Dennoch sollten Sie darauf achten, wie genau diese Voraussetzungen erfüllt sind.

So identifizieren Sie ungewöhnliche Fälle:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Daten
  Ungewöhnliche Fälle identifizieren...

Abbildung 5-1
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Variablen"

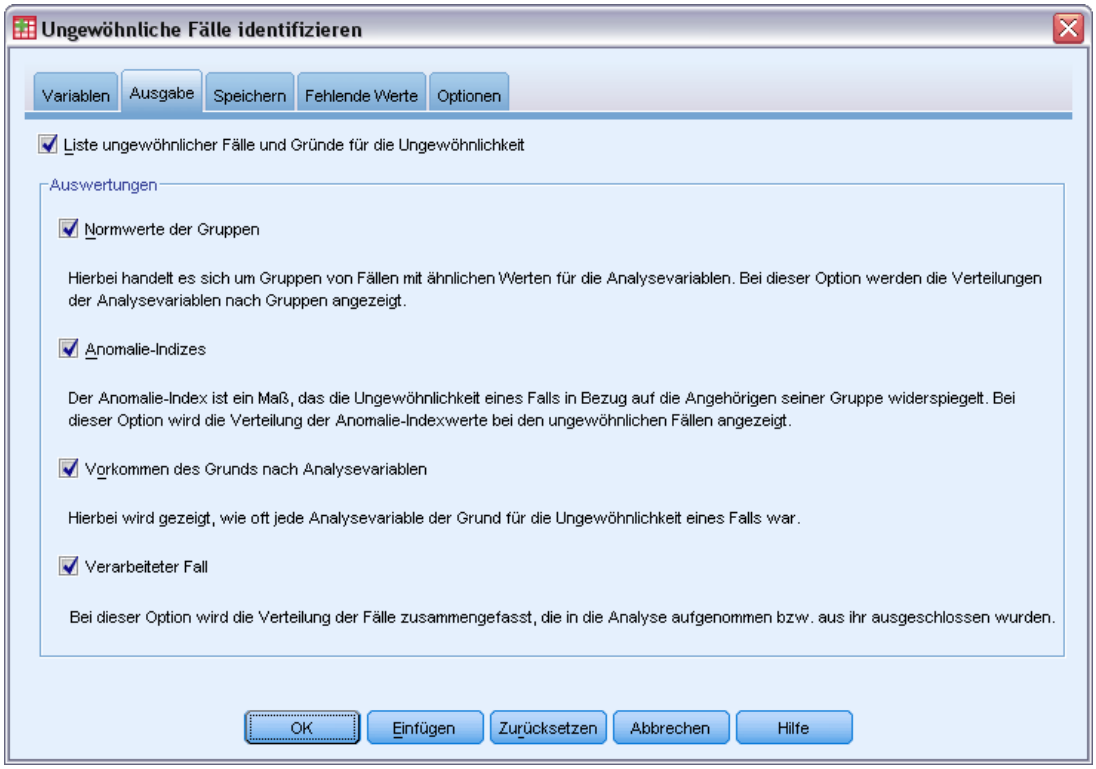


▶ Wählen Sie mindestens eine Analysevariable aus.

▶ Wahlweise können Sie eine Fallbezeichnervariable zum Beschriften der Ausgabe auswählen.

Ungewöhnliche Fälle identifizieren: Ausgabe

Abbildung 5-2
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Ausgabe"


Ungewöhnliche Fälle identifizieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Liste ungewöhnlicher Fälle und Gründe für die Ungewöhnlichkeit
Auswertungen
Normwerte der Gruppen
Hierbei handelt es sich um Gruppen von Fällen mit ähnlichen Werten für die Analysevariablen. Bei dieser Option werden die Verteilungen der Analysevariablen nach Gruppen angezeigt.
Anomalie-Indizes
Der Anomalie-Index ist ein Maß, das die Ungewöhnlichkeit eines Falls in Bezug auf die Angehörigen seiner Gruppe widerspiegelt. Bei dieser Option wird die Verteilung der Anomalie-Indexwerte bei den ungewöhnlichen Fällen angezeigt.
Vorkommen des Grunds nach Analysevariablen
Hierbei wird gezeigt, wie oft jede Analysevariable der Grund für die Ungewöhnlichkeit eines Falls war.
Verarbeiteter Fall
Bei dieser Option wird die Verteilung der Fälle zusammengefasst, die in die Analyse aufgenommen bzw. aus ihr ausgeschlossen wurden.
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

Liste ungewöhnlicher Fälle und Gründe für die Ungewöhnlichkeit. Bei dieser Option werden drei Tabellen erstellt:

- Die Liste der Indizes anomaler Fälle zeigt die als ungewöhnlich identifizierten Fälle und deren entsprechende Anomalie-Indexwerte an.
- Die Liste der Gruppen-IDs anomaler Fälle zeigt ungewöhnliche Fälle und die Informationen über deren entsprechende Gruppen an.
- Die Liste der Gründe anomaler Fälle zeigt die Fallanzahl, die Grundvariable, den Einflusswert der Variablen, den Wert der Variablen und den Normwert der Variablen für jeden Grund an.

Alle Tabellen werden nach Anomalie-Index in absteigender Reihenfolge sortiert. Darüber hinaus werden die IDs der Fälle angezeigt, wenn auf der Registerkarte "Variablen" eine Fallbezeichnervariable angegeben wurde.

Auswertung. Mit den Steuerlementen in diesem Gruppenfeld werden Auswertungen der Verteilungen erstellt.

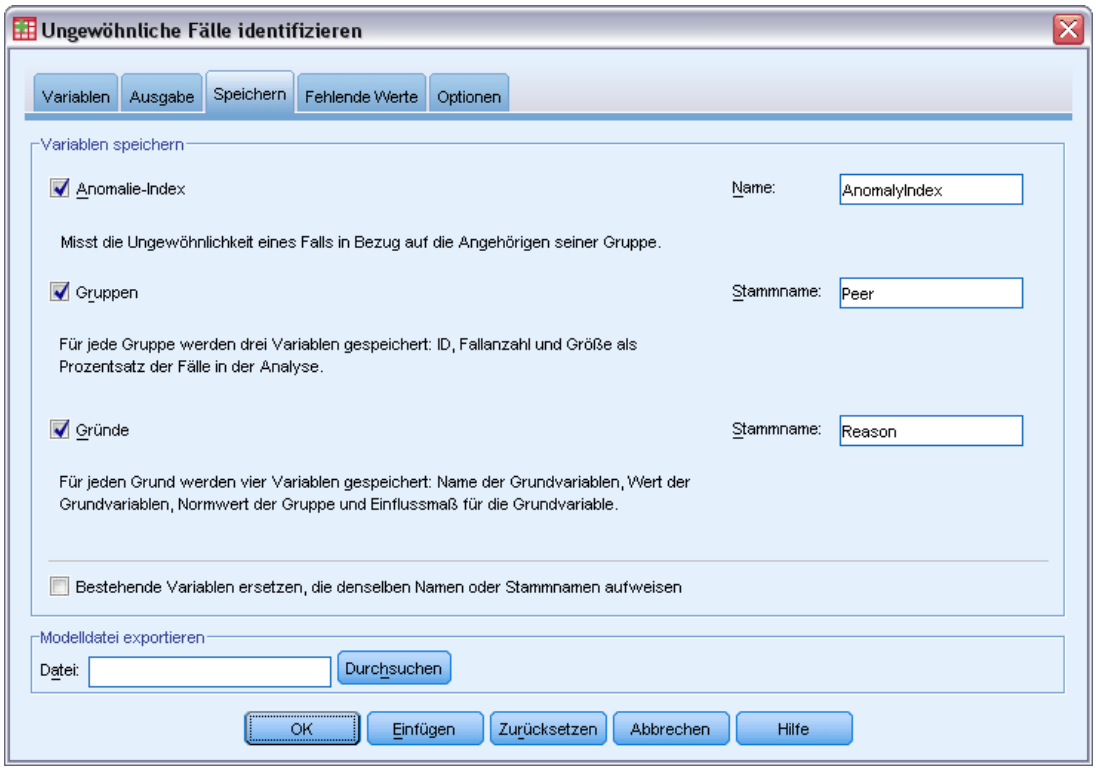
- **Normwerte der Gruppen.** Bei dieser Option wird die Tabelle für die Normwerte der stetigen Variablen (wenn die Analyse stetige Variablen umfasst) und die Tabelle für die Normwerte der kategorialen Variablen (wenn die Analyse kategoriale Variable umfasst) angezeigt. Die Tabelle für die Normwerte der stetigen Variablen enthält den Mittelwert und die Standabweichung jeder stetigen Variablen für jede Gruppe. Die Tabelle für die Normwerte

der kategorialen Variablen enthält den Modalwert (die häufigste Kategorie), die Häufigkeit und die Häufigkeit in Prozent jeder kategorialen Variablen für jede Gruppe. Der Mittelwert einer stetigen Variablen und der Modalwert einer kategorialen Variablen werden in der Analyse als Normwerte verwendet.

- **Anomalie-Indizes.** Die Auswertung des Anomalie-Index enthält deskriptive Statistiken für die Anomalie-Indizes der Fälle, die als am ungewöhnlichsten identifiziert wurden.
- **Vorkommen des Grunds nach Analysevariablen.** Die Tabelle zeigt pro Grund die Häufigkeit und die Häufigkeit in Prozent des Vorkommens jeder Variable als Grund an. Die Tabelle führt auch deskriptive Statistiken über den Einfluss jeder Variablen auf. Wenn die maximale Anzahl von Gründen auf der Registerkarte "Optionen" auf 0 festgelegt wurde, steht diese Option nicht zur Verfügung.
- **Verarbeitete Fälle.** Die Zusammenfassung der Fallverarbeitung enthält Häufigkeiten und Häufigkeiten in Prozent für alle Fälle in der Arbeitsdatei, die in die Analyse aufgenommenen und ausgeschlossenen Fälle und die Fälle in jeder Gruppe.

Ungewöhnliche Fälle identifizieren: Speichern

Abbildung 5-3
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Speichern"



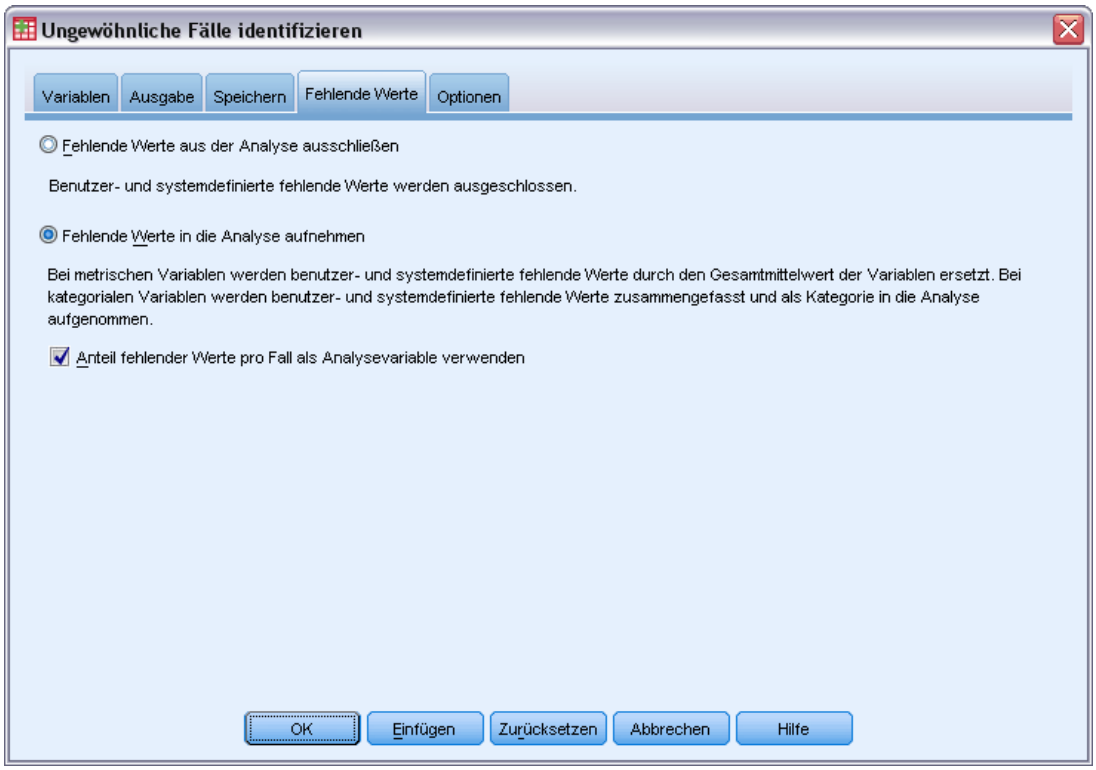
Variablen speichern. Mithilfe der Steuerelemente in diesem Gruppenfeld können Sie Modellvariablen in der Arbeitsdatei speichern. Sie können auch festlegen, dass vorhandene Variablen ersetzt werden, deren Namen mit den zu speichernden Variablen kollidieren.

- **Anomalie-Index.** Speichert für jeden Fall den Wert des Anomalie-Index in einer Variablen mit dem angegebenen Namen.
- **Gruppen.** Speichert die Gruppen-ID, die Fallanzahl und die Größe als Prozentsatz für jeden Fall in Variablen mit dem angegebenen Stammnamen. Wenn für den Stammnamen zum Beispiel *Gruppe* angegeben wurde, werden die Variablen *GruppeID*, *GruppeGröße* und *GruppePrztGröße* erzeugt. *GruppeID* stellt die Gruppen-ID des Falls dar, *GruppeGröße* die Gruppengröße und *GruppePrztGröße* die Gruppengröße als Prozentsatz.
- **Gründe.** Speichert Sets von Grundvariablen mit dem angegebenen Stammnamen. Ein Set von Grundvariablen besteht aus dem Namen einer Variablen, die einen Grund darstellt, dem Einflussmaß der Variablen, dem Variablenwert und dem Normwert. Die Anzahl der Sets hängt von der Anzahl der angeforderten Gründe ab (angegeben auf der Registerkarte "Optionen"). Wenn als Stammname zum Beispiel *Grund* angegeben wurde, werden die Variablen *GrundVar_k*, *GrundMaß_k*, *GrundWert_k* und *GrundNormwert_k* erzeugt, wobei *k* den *k*-ten Grund darstellt. Diese Option steht nicht zur Verfügung, wenn die Anzahl der Gründe auf 0 festgelegt wurde.

Modelldatei exportieren. Hiermit können Sie das Modell im XML-Format speichern.

Ungewöhnliche Fälle identifizieren: Fehlende Werte

Abbildung 5-4
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Fehlende Werte"


Ungewöhnliche Fälle identifizieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Fehlende Werte aus der Analyse ausschließen
Benutzer- und systemdefinierte fehlende Werte werden ausgeschlossen.
Fehlende Werte in die Analyse aufnehmen
Bei metrischen Variablen werden benutzer- und systemdefinierte fehlende Werte durch den Gesamtmittelwert der Variablen ersetzt. Bei kategorialen Variablen werden benutzer- und systemdefinierte fehlende Werte zusammengefasst und als Kategorie in die Analyse aufgenommen.
Anteil fehlender Werte pro Fall als Analysevariable verwenden
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

Auf der Registerkarte "Fehlende Werte" kann die Behandlung benutzerdefinierter und systemdefinierter fehlender Werte festgelegt werden.

- **Fehlende Werte aus der Analyse ausschließen.** Fälle mit fehlenden Werten werden aus der Analyse ausgeschlossen.
- **Fehlende Werte in die Analyse aufnehmen.** Fehlende Werte von stetigen Variablen werden durch deren entsprechenden Gesamtmittelwert ersetzt. Fehlende Kategorien von kategorialen Variablen werden gruppiert und als gültige Kategorie behandelt. Die verarbeiteten Variablen werden anschließend in der Analyse verwendet. Sie können die Erzeugung einer zusätzlichen Variable anfordern, die den Anteil der fehlenden Variablen in jedem Fall darstellt, und diese Variable in der Analyse verwenden.

Ungewöhnliche Fälle identifizieren: Optionen

Abbildung 5-5
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Optionen"

Kriterien zum Identifizieren ungewöhnlicher Fälle. Diese Optionen bestimmen, wie viele Fälle in die Liste der Anomalien aufgenommen werden.

- **Prozentsatz der Fälle mit den höchsten Anomalie-Indexwerten.** Geben Sie eine positive Zahl kleiner oder gleich 100 ein.
- **Feste Anzahl von Fällen mit den höchsten Anomalie-Indexwerten.** Geben Sie eine positive Ganzzahl an, die kleiner oder gleich der Gesamtzahl der in der Analyse verwendeten Fälle in der Arbeitsdatei ist.
- **Nur Fälle identifizieren, deren Anomalie-Index größer oder gleich einem Minimalwert ist.** Geben Sie eine nichtnegative Zahl an. Ein Fall wird als Anomalie betrachtet, wenn sein Anomalie-Index größer oder gleich dem angegebenen Trennwert ist. Diese Option wird

zusammen mit den Optionen Prozentsatz der Fälle und Feste Anzahl von Fällen verwendet. Wenn Sie beispielsweise eine feste Anzahl von 50 Fällen und einen Trennwert von 2 angeben, besteht die Anomalie-Liste höchstens aus 50 Fällen, von denen jeder einen Anomalie-Indexwert größer oder gleich 2 aufweist.

Anzahl von Gruppen. Die Prozedur sucht nach der besten Anzahl von Gruppen zwischen dem angegebenen Minimal- und Maximalwert. Die Werte müssen positive Ganzzahlen sein, und das Minimum darf das Maximum nicht überschreiten. Wenn die angegebenen Werte gleich sind, setzt die Prozedur eine feste Anzahl von Gruppen voraus.

Anmerkung: Abhängig von der Variation in den Daten können Situationen auftreten, in denen die Daten weniger Gruppen unterstützen können als als Minimum angegeben. In einer solchen Situation erzeugt die Prozedur eine kleinere Anzahl von Gruppen.

Maximale Anzahl von Gründen. Ein Grund besteht aus dem Variablen-Einflussmaß, dem Variablennamen für diesen Grund, dem Wert der Variablen und dem Wert der entsprechenden Gruppe. Geben Sie eine nichtnegative Ganzzahl an. Wenn dieser Wert größer oder gleich der Anzahl der verarbeiteten Variablen ist, die in der Analyse verwendet werden, werden alle Variablen angezeigt.

Zusätzliche Funktionen beim Befehl DETECTANOMALY

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Sie können einige Variablen in der Arbeitsdatei aus der Analyse ausschließen, ohne dass ausdrücklich alle Analysevariablen angegebenen werden müssen (mit dem Unterbefehl `EXCEPT`).
- Sie können eine Korrektur angeben, um den Einfluss von stetigen und kategorialen Variablen auszutarieren (mit dem Schlüsselwort `MLWEIGHT` im Unterbefehl `CRITERIA`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
6

Optimales Klassieren

Die Prozedur “Optimales Klassieren” diskretisiert eine oder mehrere metrische Variablen (im Folgenden als **Klassierungs-Eingabevariablen** (Binning-Eingabevariablen) bezeichnet), indem die Werte der einzelnen Variablen auf verschiedene Klassen verteilt werden. Die Klassenbildung ist in Bezug auf eine kategoriale Führungsvariable optimal, die den Klassierungsvorgang “überwacht”. Anstatt der ursprünglichen Datenwerte können dann die Klassen zur weiteren Analyse verwendet werden.

Beispiele. Für die Verringerung der unterschiedlichen Werte, die eine Variable annehmen kann, gibt es verschiedenen Anwendungsmöglichkeiten. Hier einige Beispiele:

- Anforderungen anderer Prozeduren an die Daten. Diskretisierte Variablen können für die Verwendung in Prozeduren, bei denen kategoriale Variablen erforderlich sind, als kategorial behandelt werden. Beispielsweise müssen für die Prozedur “Kreuztabellen” alle Variablen kategorial sein.
- Datenschutz. Die Angabe von gebinnten Werten anstelle der tatsächlichen Werte in Berichten kann zur Gewährleistung des Datenschutzes bei Ihren Datenquellen beitragen. Die Prozedur “Optimales Binning” kann eine Orientierung für die Auswahl der Klassen bieten.
- Schnellere Durchführung. Einige Prozeduren sind effizienter, wenn sie mit einer reduzierten Anzahl an unterschiedlichen Werten arbeiten. So lässt sich beispielsweise die Geschwindigkeit der multinomialen logistischen Regression durch die Verwendung diskretisierter Variablen erhöhen.
- Ermittlung vollständiger oder quasi vollständiger Datentrennung.

Optimales Binning im Vergleich zum visuellen Binning In den Dialogfeldern von “Visuelles Binning” stehen Ihnen mehrere automatische Methoden zur Erstellung von Klassen ohne die Verwendung einer Führungsvariablen zur Verfügung. Diese Regeln für unüberwachtes Binning sind nützlich für die Erstellung deskriptiver Statistiken, wie beispielsweise Häufigkeitstabellen, “Optimales Binning” ist am besten, wenn das Endziel in der Erstellung eines Vorhersagemodells besteht.

Ausgabe. Mit dieser Prozedur werden Tabellen mit Trennwerten für die Klassen und deskriptive Statistiken für jede Klassierungs-Eingabevariable erstellt. Zusätzlich können Sie neue Variablen im aktiven Daten-Set speichern, die die klassierten Werte der Klassierungs-Eingabevariablen enthalten und die Klassierungsregeln als Befehlssyntax zur Verwendung bei der Diskretisierung neuer Daten speichern.

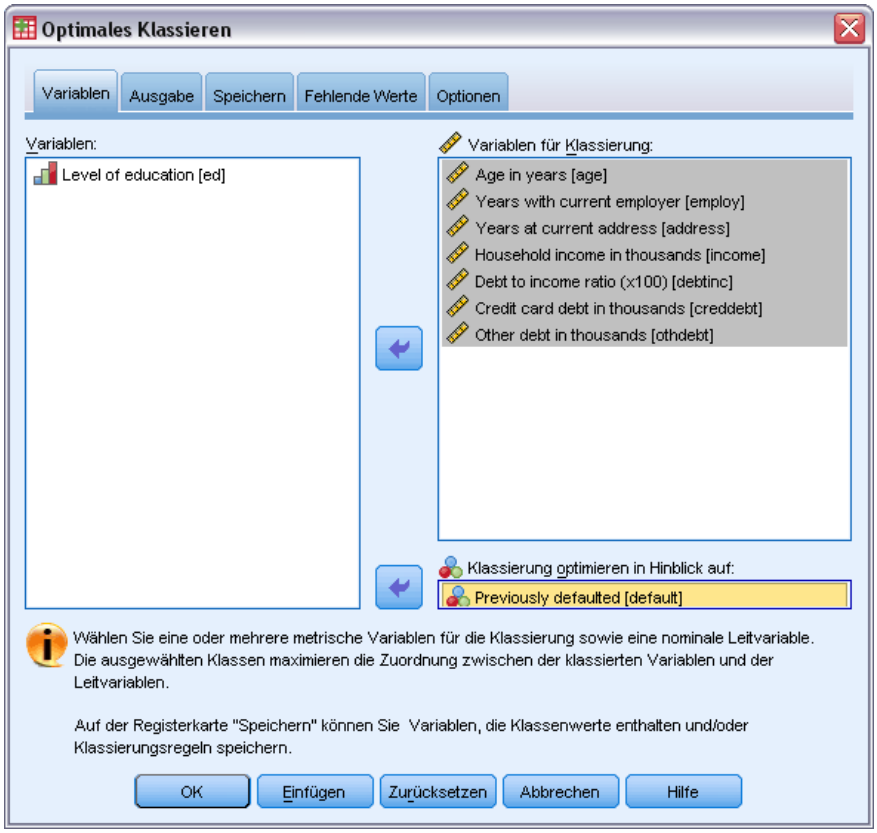
Daten. Bei dieser Prozedur wird davon ausgegangen, dass es sich bei den Binning-Eingabevariablen um metrische, numerische Variablen handelt. Die Führungsvariable sollte kategorial sein. Es kann sich dabei um eine String-Variable oder eine numerische Variable handeln.

So erhalten Sie ein optimales Binning:

Wählen Sie die folgenden Befehle aus den Menüs aus:

Transformieren
Optimales Klassieren...

Abbildung 6-1
Dialogfeld "Optimales Klassieren", Registerkarte "Variablen"

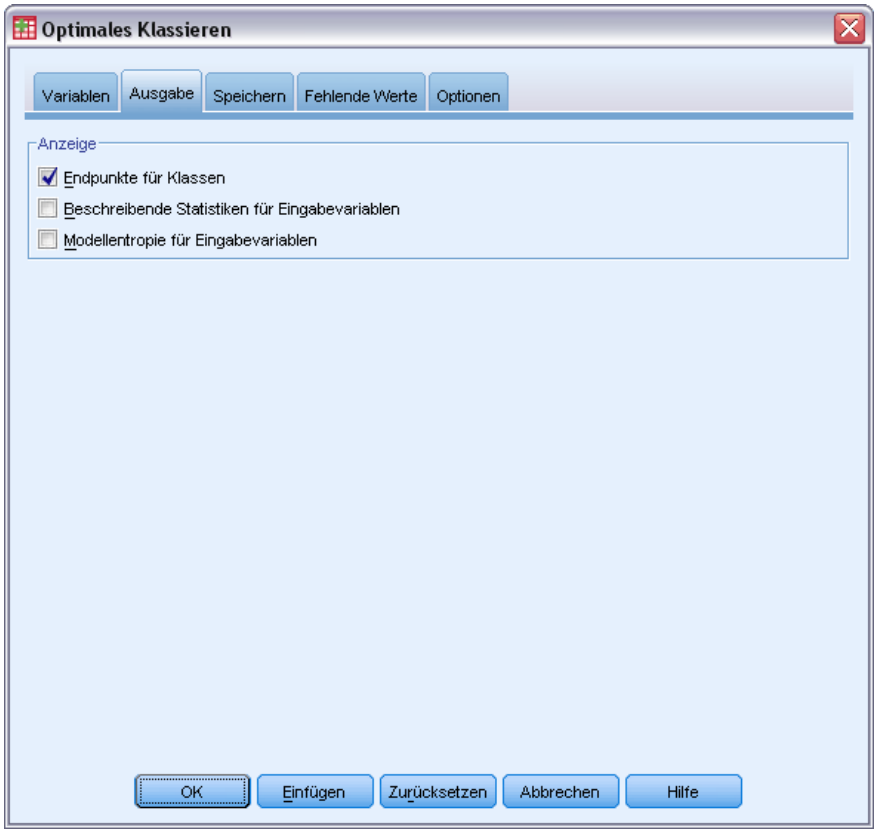

Optimales Klassieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Variablen:
Level of education [ed]
Variablen für Klassierung:
Age in years [age]
Years with current employer [employ]
Years at current address [address]
Household income in thousands [income]
Debt to income ratio (x100) [debtinc]
Credit card debt in thousands [creddebt]
Other debt in thousands [othdebt]
Klassierung optimieren in Hinblick auf:
Previously defaulted [default]
Wählen Sie eine oder mehrere metrische Variablen für die Klassierung sowie eine nominale Leitvariable. Die ausgewählten Klassen maximieren die Zuordnung zwischen der klassierten Variablen und der Leitvariablen.
Auf der Registerkarte "Speichern" können Sie Variablen, die Klassenwerte enthalten und/oder Klassierungsregeln speichern.
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

- Wählen Sie mindestens eine Binning-Eingabevariable aus.
- Wählen Sie eine Führungsvariable aus.

Variablen, die die klassierten Datenwerte enthalten, werden nicht standardmäßig erstellt. Auf der Registerkarte Speichern können Sie diese Variablen speichern.

Optimales Binning – Ausgabe

Abbildung 6-2
Dialogfeld "Optimales Klassieren", Registerkarte "Ausgabe"

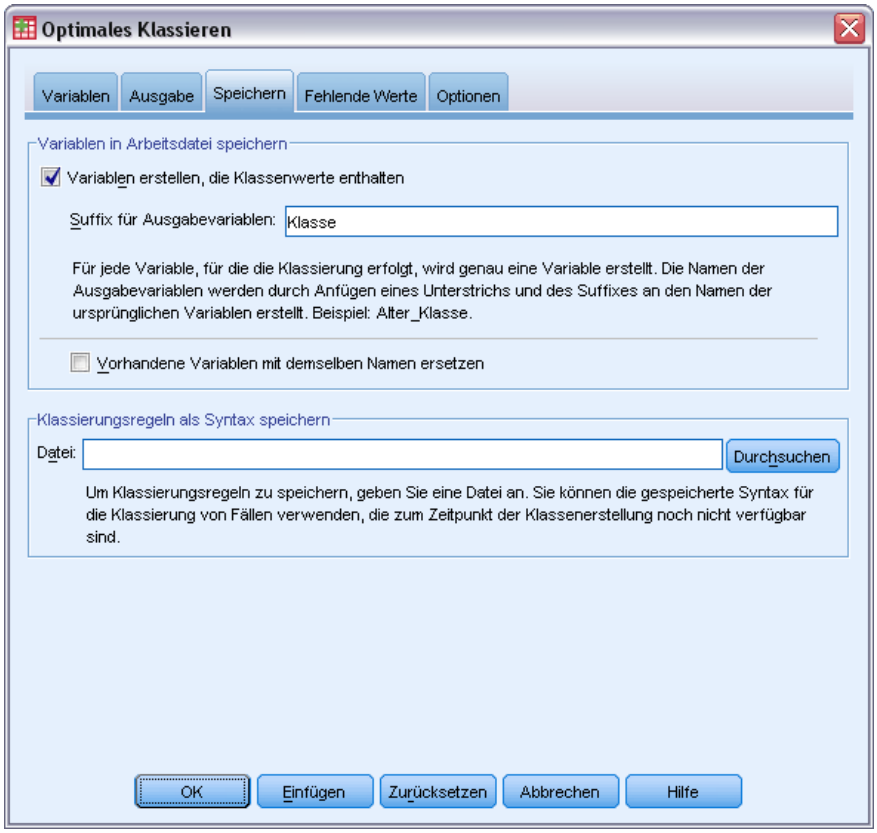


Die Registerkarte "Ausgabe" steuert die Anzeige der Ergebnisse.

- **Endpunkte für Klassen.** Zeigt das Set an Endpunkten für die einzelnen Klassierungs-Eingabevariablen an.
- **Beschreibende Statistiken für Binning-Variablen.** Diese Option zeigt für die einzelnen Binning-Eingabevariablen die Anzahl der Fälle mit gültigen Werten, die Anzahl der Fälle mit fehlenden Werten, die Anzahl der verschiedenen gültigen Werte sowie die Minimal- und Maximalwerte an. Für die Führungsvariable zeigt diese Option die Klassenverteilung für alle zugehörigen Binning-Eingabevariaben an.
- **Modellentropie für Binning-Variable.** Für jede Binning-Eingabevariable zeigt diese Option ein Maß für die Vorhersagegenauigkeit der Variablen hinsichtlich der Führungsvariablen an.

Optimales Binning – Speichern

Abbildung 6-3
Dialogfeld "Optimales Klassieren", Registerkarte "Speichern"

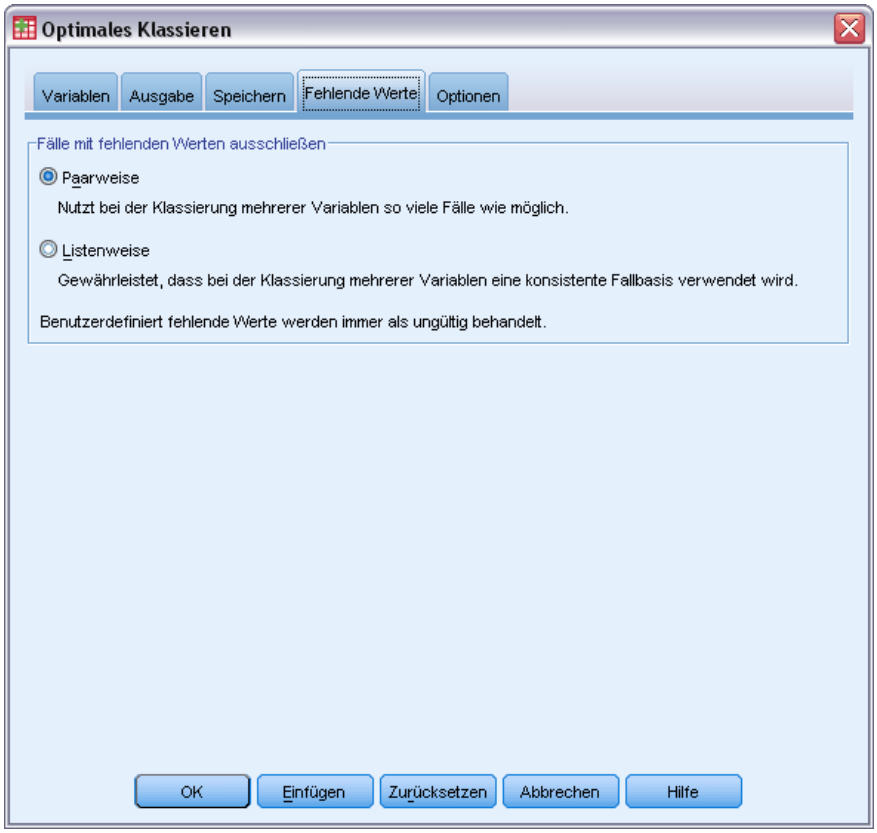

Optimales Klassieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Variablen in Arbeitsdatei speichern
Variablen erstellen, die Klassenwerte enthalten
Suffix für Ausgabevariablen:
Klasse
Für jede Variable, für die die Klassierung erfolgt, wird genau eine Variable erstellt. Die Namen der Ausgabevariablen werden durch Anfügen eines Unterstrichs und des Suffixes an den Namen der ursprünglichen Variablen erstellt. Beispiel: Alter_Klasse.
Vorhandene Variablen mit demselben Namen ersetzen
Klassierungsregeln als Syntax speichern
Datei:
Durchsuchen
Um Klassierungsregeln zu speichern, geben Sie eine Datei an. Sie können die gespeicherte Syntax für die Klassierung von Fällen verwenden, die zum Zeitpunkt der Klassenerstellung noch nicht verfügbar sind.
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

Variablen in Arbeitsdatei speichern. In der weiteren Analyse können anstelle der ursprünglichen Variablen Variablen verwendet werden, die die gebinnten Datenwerte enthalten.

Klassierungsregeln als Syntax speichern. Generiert Befehlssyntax, die für die Klassierung von anderen Daten-Sets verwendet werden kann. Die Umkodierungsregeln beruhen auf den vom Klassierungsalgorithmus bestimmten Trennwerten.

Optimales Binning – Fehlende Werte

Abbildung 6-4
Dialogfeld "Optimales Binning", Registerkarte "Fehlende Werte"

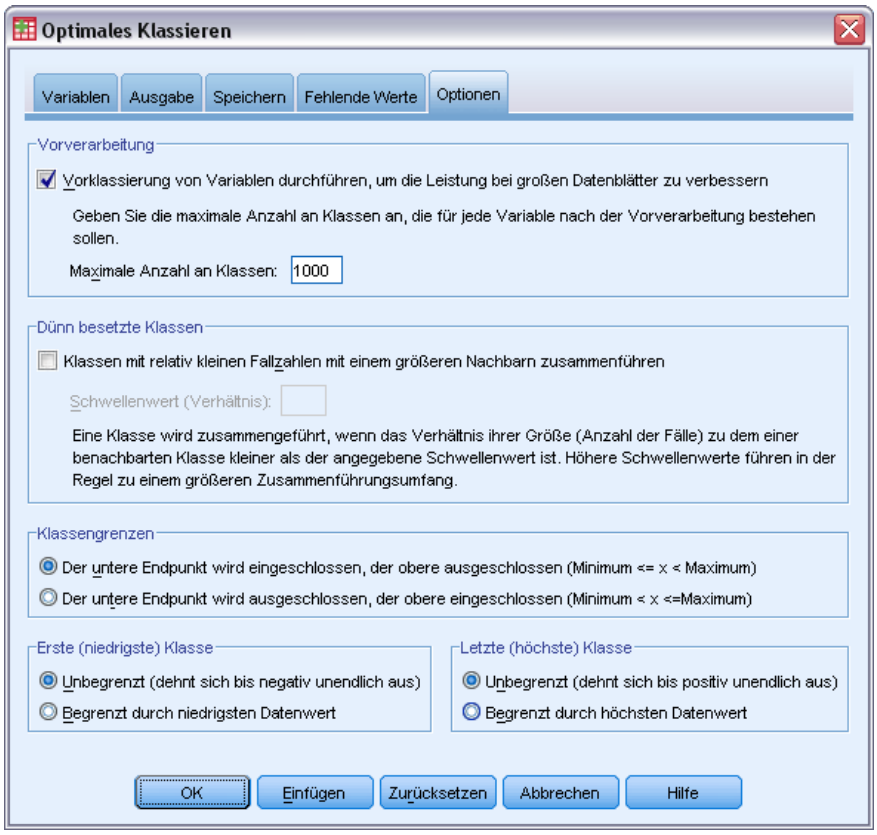


Auf der Registerkarte "Fehlende Werte" wird angegeben, ob der Umgang mit fehlenden Werten anhand eines listenweisen oder paarweisen Ausschlusses erfolgt. Benutzerdefinierte fehlende Werte werden stets als ungültig behandelt. Bei der Umkodierung der ursprünglichen Variablenwerte in eine neue Variable werden benutzerdefiniert fehlende Werte in systemdefiniert fehlende Werte umgewandelt.

- **Paarweise.** Diese Option operiert auf der Basis der einzelnen Paare aus Führungsvariabler und Binning-Eingabevariabler. Die Prozedur verwendet alle Fälle mit nichtfehlenden Werten bei der Führungs- und Binning-Eingabevariablen.
- **Listenweise** Diese Option wird auf alle auf der Registerkarte "Variablen" angegebenen Variablen angewendet. Wenn bei einem Fall eine Variable fehlt, wird der gesamte Fall ausgeschlossen.

Optimales Binning – Optionen

Abbildung 6-5
Dialogfeld "Optimales Binning", Registerkarte "Optionen"


Optimales Klassieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Vorverarbeitung
Vorklassierung von Variablen durchführen, um die Leistung bei großen Datenblätter zu verbessern
Geben Sie die maximale Anzahl an Klassen an, die für jede Variable nach der Vorverarbeitung bestehen sollen.
Maximale Anzahl an Klassen:
1000
Dünn besetzte Klassen
Klassen mit relativ kleinen Fallzahlen mit einem größeren Nachbarn zusammenführen
Schwellenwert (Verhältnis):
Eine Klasse wird zusammengeführt, wenn das Verhältnis ihrer Größe (Anzahl der Fälle) zu dem einer benachbarten Klasse kleiner als der angegebene Schwellenwert ist. Höhere Schwellenwerte führen in der Regel zu einem größeren Zusammenführungsumfang.
Klassengrenzen
Der untere Endpunkt wird eingeschlossen, der obere ausgeschlossen (Minimum <= x < Maximum)
Der untere Endpunkt wird ausgeschlossen, der obere eingeschlossen (Minimum < x <=Maximum)
Erste (niedrigste) Klasse
Unbegrenzt (dehnt sich bis negativ unendlich aus)
Begrenzt durch niedrigsten Datenwert
Letzte (höchste) Klasse
Unbegrenzt (dehnt sich bis positiv unendlich aus)
Begrenzt durch höchsten Datenwert
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

Vorverarbeitung. Das "Pre-Binning" von Binning-Eingabevariablen mit vielen verschiedenen Werten kann die Verarbeitung ohne größere Qualitätseinbußen bei den endgültigen Klassen beschleunigen. Der Wert für die maximale Anzahl an Klassen stellt lediglich die Obergrenze für die Anzahl der erstellten Klassen dar. Wenn Sie also 1000 als Maximalwert angeben, eine Binning-Eingabevariable jedoch weniger als 1000 verschiedene Werte aufweist, werden so viele vorverarbeitete Klasen für die Binning-Eingabevariable erstellt wie verschiedene Klassen in der Binning-Eingabevariablen enthalten sind.

Dünn besetzte Klassen. Gelegentlich kann die Prozedur zu Klassen mit sehr wenigen Fällen führen. Mit der folgenden Strategie können diese Pseudotrennwerte gelöscht werden:

- Angenommen, der Algorithmus hat für eine Variable $n_{\text{endgültig}}$ Trennwerte und daher $n_{\text{endgültig}}+1$ Klassen gefunden. Für die Klassen $i = 2, ..., n_{\text{endgültig}}$ (von der Klasse mit dem zweitniedrigsten Wert bis zur Klasse mit dem zweithöchsten Wert) wird Folgendes berechnet:

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}),sizeof(b_{i+1}))}$$

Dabei ist *sizeof(b)* die Anzahl der Fälle in der Klasse.

- Wenn dieser Wert kleiner ist als der angegebene Zusammenführungsschwellenwert, dann wird b_i als dünn besetzt betrachtet und mit b_{i-1} oder b_{i+1} zusammengeführt, je nachdem, welche Klasse die niedrigere Klasseninformationsentropie aufweist.

  Bei dieser Prozedur wird ein einzelner Durchlauf durch die Klassen vorgenommen.

  Binning von Endpunkten. Bei dieser Option wird angegeben, wie die Untergrenze eines Intervalls festgelegt wird. Da die Prozedur die Trennwerte automatisch ermittelt, ist dies weitgehend eine Frage der Vorlieben.

  Erste (niedrigste) Klasse/Letzte (höchste) Klasse. Diese Optionen geben an, wie die minimalen und maximalen Trennwerte für die einzelnen Klassierungs-Eingabevariablen festgelegt werden. Im Allgemeinen geht die Prozedur davon aus, dass die Binning-Eingabevariablen einen beliebigen Wert der reellen Zahlen annehmen können, aber wenn es theoretische oder praktische Gründe für die Begrenzung des Bereichs gibt, können Sie den gewünschten niedrigsten und/oder höchsten Wert angeben.

Zusätzliche Funktionen beim Befehl OPTIMAL BINNING

Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Sie können mithilfe der Methode der gleichen Häufigkeiten unüberwachtes Binning durchführen (mit dem Unterbefehl `CRITERIA`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Teil II: Beispiele

Kapitel 7

Daten validieren

Mit der Prozedur “Daten validieren” können verdächtige und ungültige Fälle, Variablen und Datenwerte identifiziert werden.

Validieren einer medizinischen Datenbank

Eine bei einem Unternehmen in der Pharmabranche angestellte Analytikerin hat die Aufgabe, die Qualität der Informationen in einem System zu überwachen. Dabei muss sie die Werte und Variablen prüfen und einen Bericht für den Leiter des Datenerfassungsteams erstellen.

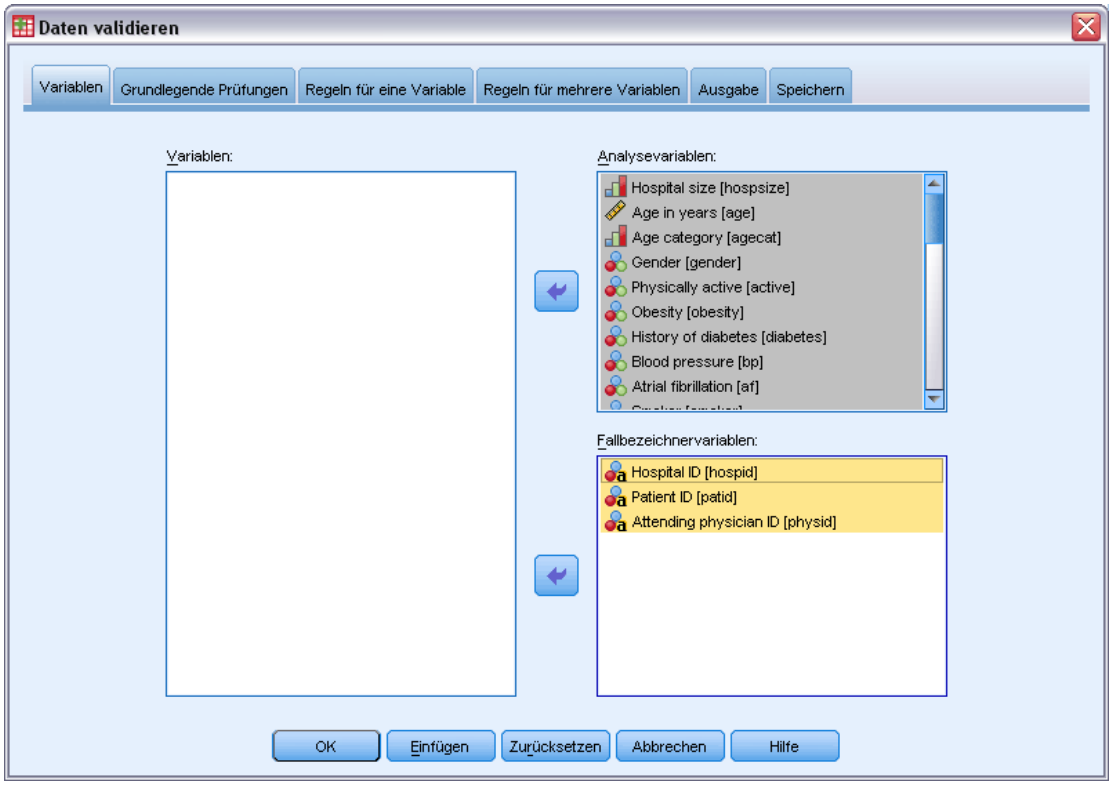
Den aktuellen Zustand der Datenbank finden Sie in der Datei *stroke_invalid.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 139. Verwenden Sie die Prozedur “Daten validieren”, um die für den Bericht benötigten Informationen zusammenzustellen. Syntax, mit denen Sie diese Analysen nachvollziehen können, befindet sich in der Datei *validatedata_stroke.sps*.

Durchführen von grundlegenden Prüfungen

▶ Um die Daten zu validieren, wählen Sie die folgenden Befehle aus den Menüs aus:

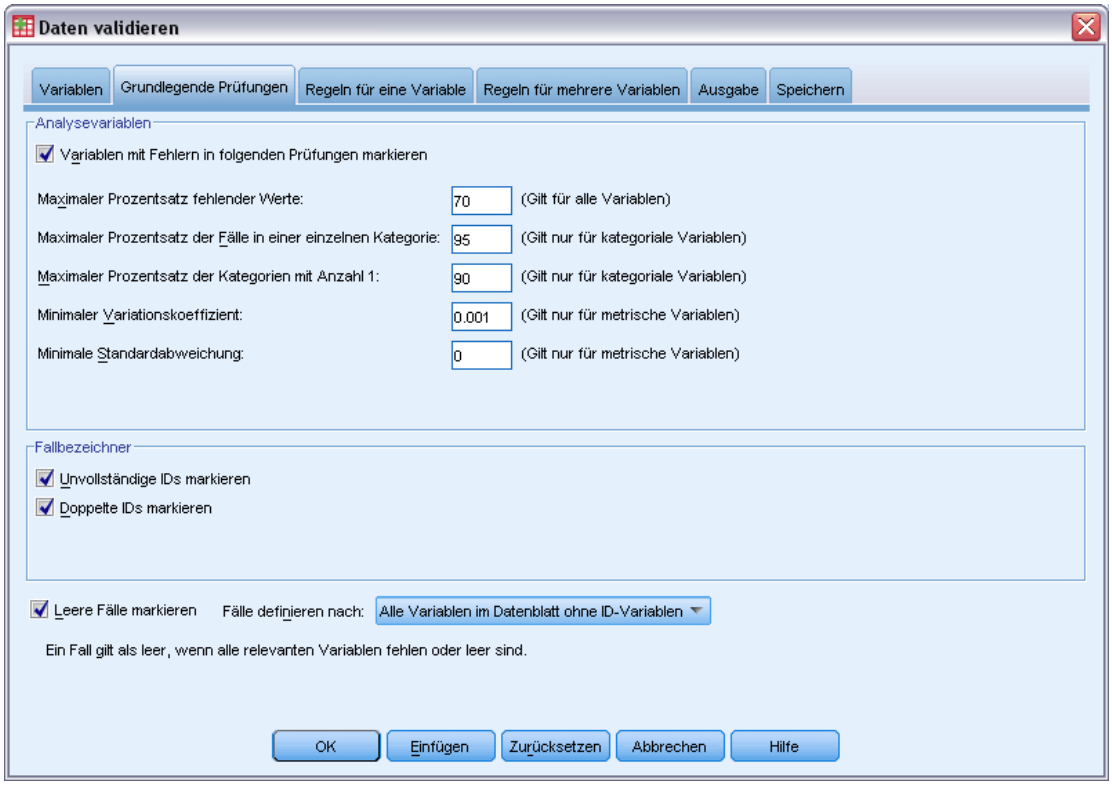
Daten
  Validierung
    Daten validieren...

Abbildung 7-1
Dialogfeld "Daten validieren", Registerkarte "Variablen"



- Wählen Sie *Hospital size* sowie die Variablen von *Age in years* bis *Recoded Barthel index at 6 months* als Analysevariablen aus.
- Wählen Sie *Hospital ID*, *Patient ID* und *Attending physician ID* als Fallbezeichnervariablen aus.
- Klicken Sie auf die Registerkarte Grundlegende Prüfungen.

Abbildung 7-2
Dialogfeld "Daten validieren", Registerkarte "Grundlegende Prüfungen"



Sie können mit den Standardeinstellungen fortfahren.

► Klicken Sie auf OK.

Warnungen

Abbildung 7-3
Warnungen

Einige oder alle der angeforderten Ausgaben werden nicht gezeigt, weil alle Fälle, Variablen oder Datenwerte die angeforderten Prüfungen bestanden haben.

Die Analysevariablen haben die grundlegenden Prüfungen bestanden, und es liegen keine leeren Fälle vor. Deshalb wird eine Warnung ausgegeben, die erläutert, warum für die grundlegenden Prüfungen keine Ausgabe vorhanden ist.

Unvollständige Identifizierung

Abbildung 7-4
Unvollständige Fallbezeichner

	Identifizierung		
Fall	Hospital ID	Patient ID	Attending physician ID
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

Wenn in den Fallbezeichnervariablen fehlende Werte vorliegen, können die entsprechenden Fälle nicht ordnungsgemäß identifiziert werden. In der vorliegenden Datendatei fehlt der Wert von *Patient ID* in Fall 288 und in den Fällen 573 und 774 sind keine Werte für *Hospital ID* vorhanden.

Gleiche Identifizierung

Abbildung 7-5
Gleiche Fallbezeichner (gezeigt werden die ersten 11)

			Identifizierung		
Gruppe mit gleicher Identifizierung	Anzahl Duplikate	Fälle mit gleicher Identifizierung	Hospital ID	Patient ID	Attending physician ID
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101, 102	QWS	5856145337	817329
11	3	104, 105, 106	QWS	1543897849	817329

Ein Fall muss eindeutig durch eine Kombination der Werte der Fallbezeichnervariablen identifiziert werden können. Hier werden die ersten 11 Einträge in der Tabelle der Fälle mit gleicher Identifizierung gezeigt. Bei diesen Duplikaten handelt es sich um Patienten, bei denen mehrere Ereignisse aufgezeichnet wurden, die für jedes Ereignis als separater Fall erfasst wurden. Da diese Informationen jeweils in einer Zeile zusammengefasst werden können, sollten diese Fälle bereinigt werden.

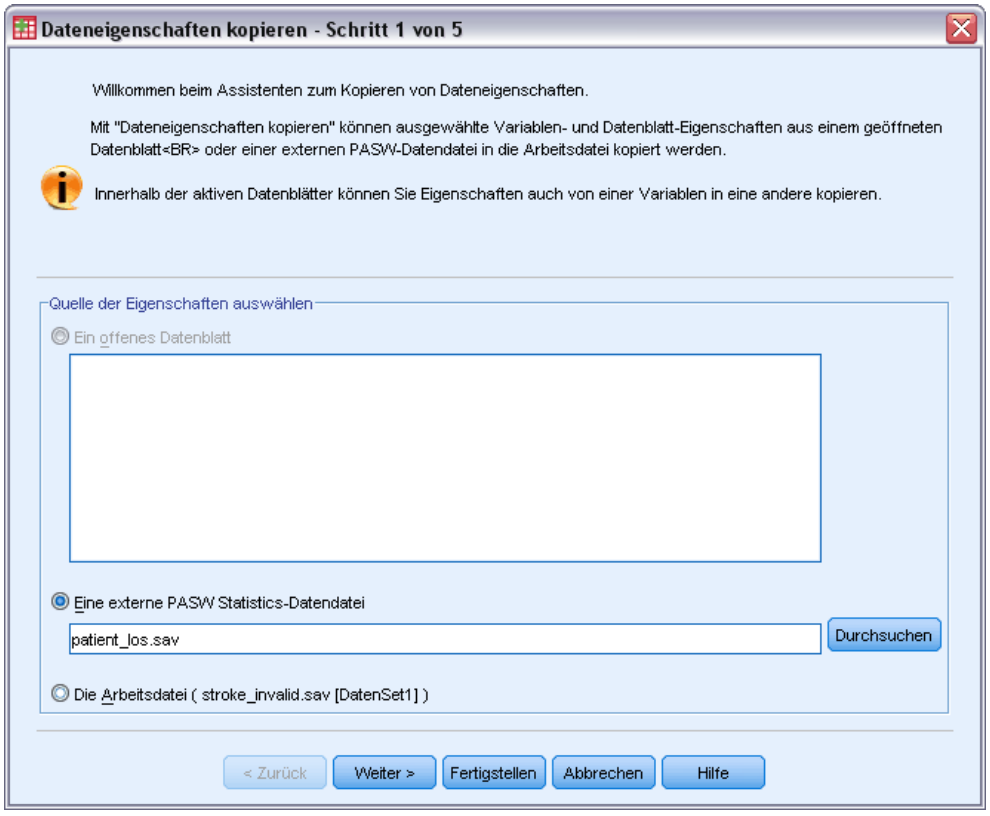
Kopieren und Verwenden von Regeln aus einer anderen Datei

Der Analytikerin fällt auf, dass die Variablen in der vorliegenden Datendatei den Variablen aus einem anderen Projekt ähneln. Die Validierungsregeln dieses Projekts wurden als Eigenschaften der entsprechenden Datendatei gespeichert und können auf die vorliegende Datendatei angewendet werden, indem die Dateneigenschaften der Datei kopiert werden.

▶ Um die Regeln aus einer anderen Datei zu kopieren, wählen Sie die folgenden Befehle aus den Menüs aus:

Daten
Dateneigenschaften kopieren...

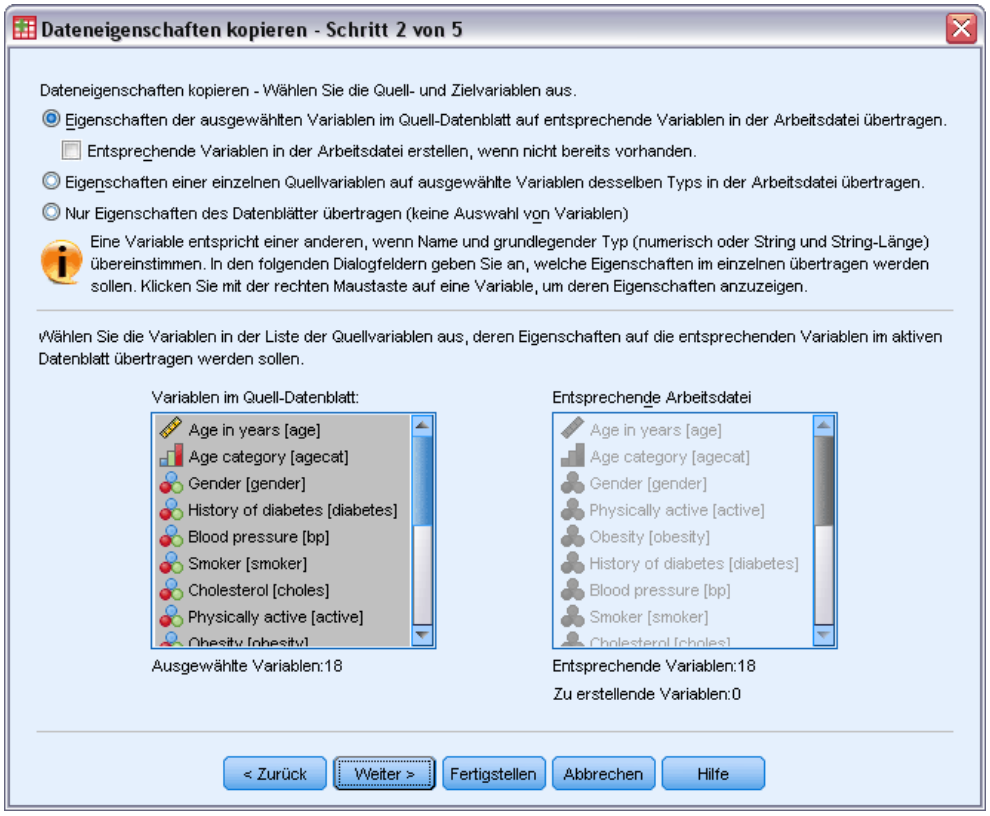
Abbildung 7-6
Kopieren von Dateneigenschaften – Schritt 1 (Begrüßung)


Dateneigenschaften kopieren - Schritt 1 von 5
Willkommen beim Assistenten zum Kopieren von Dateneigenschaften.
Mit "Dateneigenschaften kopieren" können ausgewählte Variablen- und Datenblatt-Eigenschaften aus einem geöffneten Datenblatt
 oder einer externen PASW-Datendatei in die Arbeitsdatei kopiert werden.
Innerhalb der aktiven Datenblätter können Sie Eigenschaften auch von einer Variablen in eine andere kopieren.
Quelle der Eigenschaften auswählen
Ein offenes Datenblatt
Eine externe PASW Statistics-Datendatei
patient_los.sav
Durchsuchen
Die Arbeitsdatei (stroke_invalid.sav [DatenSet1])
< Zurück
Weiter >
Fertigstellen
Abbrechen
Hilfe

▶ Wählen Sie aus, dass die Eigenschaften aus einer externen PASW Statistics-Datendatei, *patient_los.sav*, kopiert werden sollen. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 139.

▶ Klicken Sie auf Weiter.

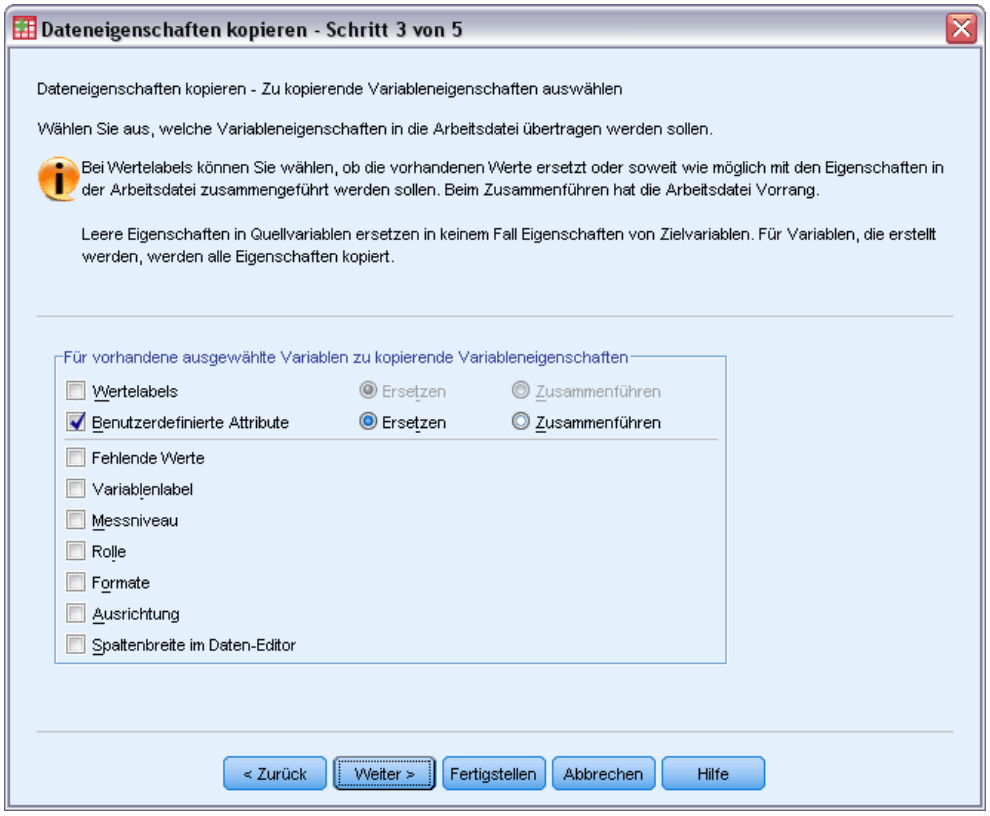
Abbildung 7-7
Kopieren von Dateneigenschaften – Schritt 2 (Variablen auswählen)


Dateneigenschaften kopieren - Schritt 2 von 5
Dateneigenschaften kopieren - Wählen Sie die Quell- und Zielvariablen aus.
Eigenschaften der ausgewählten Variablen im Quell-Datenblatt auf entsprechende Variablen in der Arbeitsdatei übertragen.
Entsprechende Variablen in der Arbeitsdatei erstellen, wenn nicht bereits vorhanden.
Eigenschaften einer einzelnen Quellvariablen auf ausgewählte Variablen desselben Typs in der Arbeitsdatei übertragen.
Nur Eigenschaften des Datenblätter übertragen (keine Auswahl von Variablen)
Eine Variable entspricht einer anderen, wenn Name und grundlegender Typ (numerisch oder String und String-Länge) übereinstimmen. In den folgenden Dialogfeldern geben Sie an, welche Eigenschaften im einzelnen übertragen werden sollen. Klicken Sie mit der rechten Maustaste auf eine Variable, um deren Eigenschaften anzuzeigen.
Wählen Sie die Variablen in der Liste der Quellvariablen aus, deren Eigenschaften auf die entsprechenden Variablen im aktiven Datenblatt übertragen werden sollen.
Variablen im Quell-Datenblatt:
Age in years [age]
Age category [agecat]
Gender [gender]
History of diabetes [diabetes]
Blood pressure [bp]
Smoker [smoker]
Cholesterol [choles]
Physically active [active]
Obesity [obesity]
Ausgewählte Variablen:18
Entsprechende Arbeitsdatei
Age in years [age]
Age category [agecat]
Gender [gender]
Physically active [active]
Obesity [obesity]
History of diabetes [diabetes]
Blood pressure [bp]
Smoker [smoker]
Cholesterol [choles]
Entsprechende Variablen:18
Zu erstellende Variablen:0
< Zurück
Weiter >
Fertigstellen
Abbrechen
Hilfe

Dies sind die Variablen aus *patient_los.sav*, deren Eigenschaften Sie in die entsprechenden Variablen in *stroke_invalid.sav* kopieren möchten.

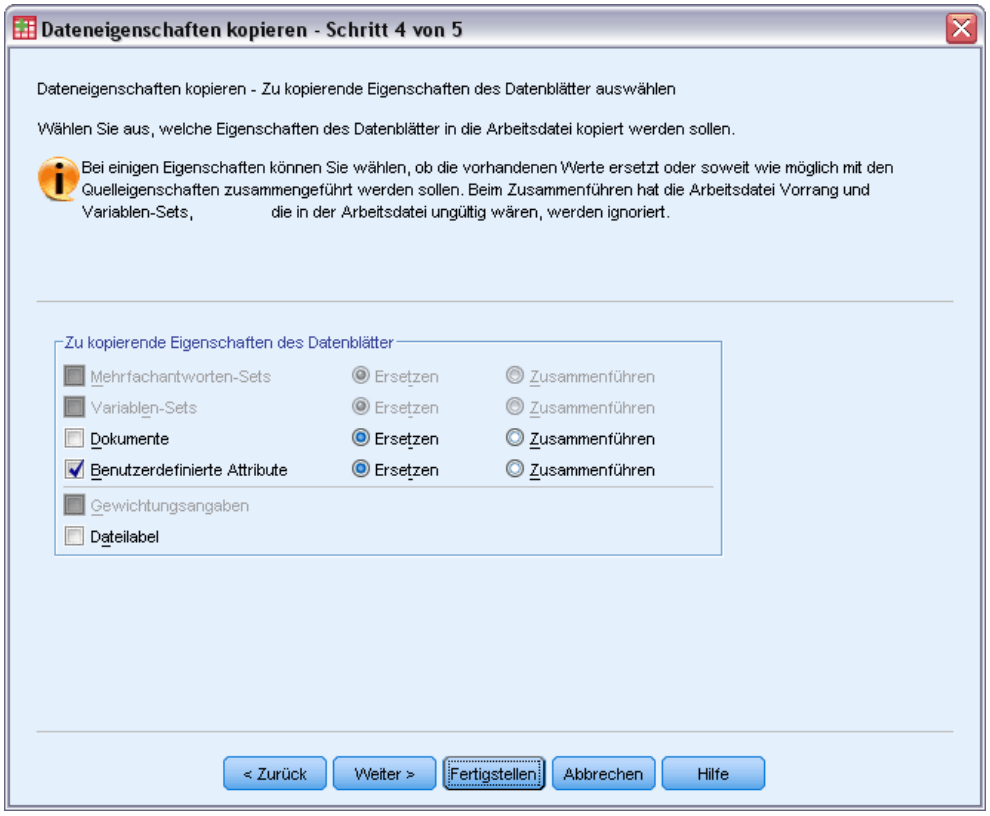
▶ Klicken Sie auf Weiter.

Abbildung 7-8
Kopieren von Dateneigenschaften – Schritt 3 (Variableneigenschaften auswählen)


Dateneigenschaften kopieren - Schritt 3 von 5
Dateneigenschaften kopieren - Zu kopierende Variableneigenschaften auswählen
Wählen Sie aus, welche Variableneigenschaften in die Arbeitsdatei übertragen werden sollen.
Bei Wertelabels können Sie wählen, ob die vorhandenen Werte ersetzt oder soweit wie möglich mit den Eigenschaften in der Arbeitsdatei zusammengeführt werden sollen. Beim Zusammenführen hat die Arbeitsdatei Vorrang.
Leere Eigenschaften in Quellvariablen ersetzen in keinem Fall Eigenschaften von Zielvariablen. Für Variablen, die erstellt werden, werden alle Eigenschaften kopiert.
Für vorhandene ausgewählte Variablen zu kopierende Variableneigenschaften
Wertelabels
Ersetzen
Zusammenführen
Benutzerdefinierte Attribute
Ersetzen
Zusammenführen
Fehlende Werte
Variablenlabel
Messniveau
Rolle
Formate
Ausrichtung
Spaltenbreite im Daten-Editor
< Zurück
Weiter >
Fertigstellen
Abbrechen
Hilfe

- Heben Sie die Auswahl aller Eigenschaften mit Ausnahme von Benutzerdefinierte Attribute auf.
- Klicken Sie auf Weiter.

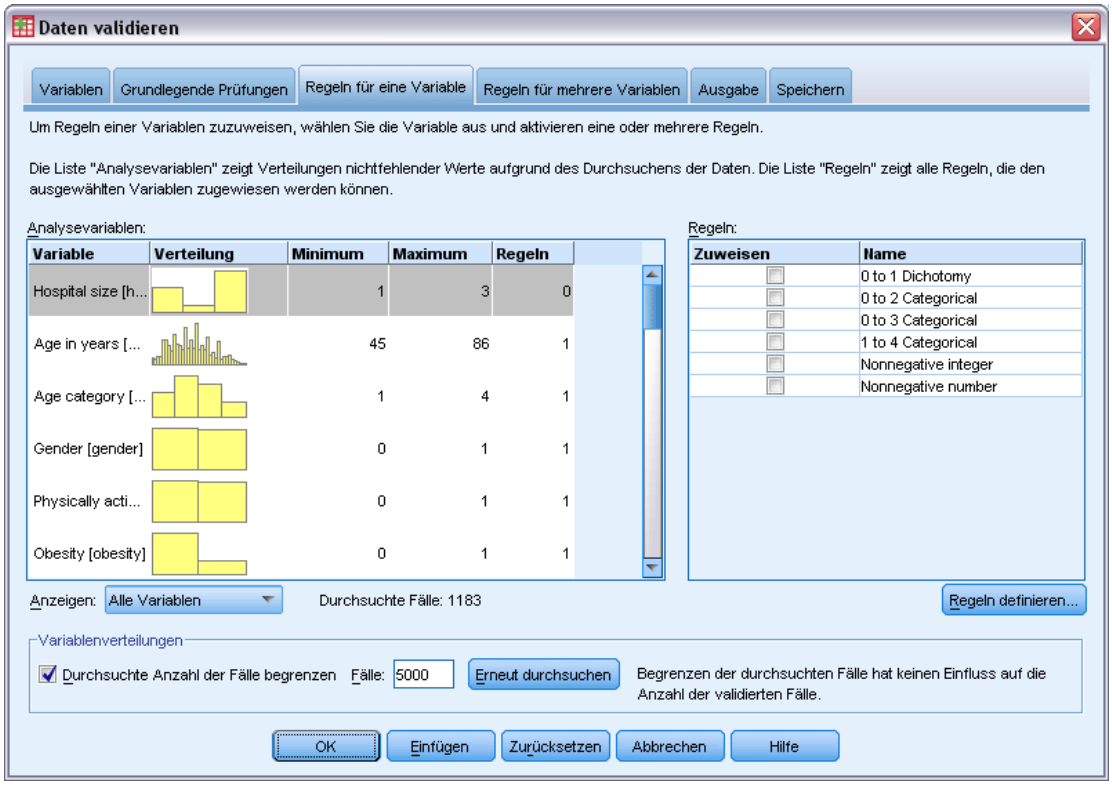
Abbildung 7-9
Kopieren von Dateneigenschaften – Schritt 4 (Daten-Set-Eigenschaften auswählen)


Dateneigenschaften kopieren - Schritt 4 von 5
Dateneigenschaften kopieren - Zu kopierende Eigenschaften des Datenblätter auswählen
Wählen Sie aus, welche Eigenschaften des Datenblätter in die Arbeitsdatei kopiert werden sollen.
Bei einigen Eigenschaften können Sie wählen, ob die vorhandenen Werte ersetzt oder soweit wie möglich mit den Quelleigenschaften zusammengeführt werden sollen. Beim Zusammenführen hat die Arbeitsdatei Vorrang und Variablen-Sets, die in der Arbeitsdatei ungültig wären, werden ignoriert.
Zu kopierende Eigenschaften des Datenblätter
Mehrfachantworten-Sets
Ersetzen
Zusammenführen
Variablen-Sets
Ersetzen
Zusammenführen
Dokumente
Ersetzen
Zusammenführen
Benutzerdefinierte Attribute
Ersetzen
Zusammenführen
Gewichtungsangaben
Dateilabel
< Zurück
Weiter >
Fertigstellen
Abbrechen
Hilfe

- Wählen Sie Benutzerdefinierte Attribute aus.
- Klicken Sie auf Fertig stellen.

Nun können Sie die Validierungsregeln verwenden.

Abbildung 7-10
Dialogfeld "Daten validieren", Registerkarte "Regeln für eine Variable"



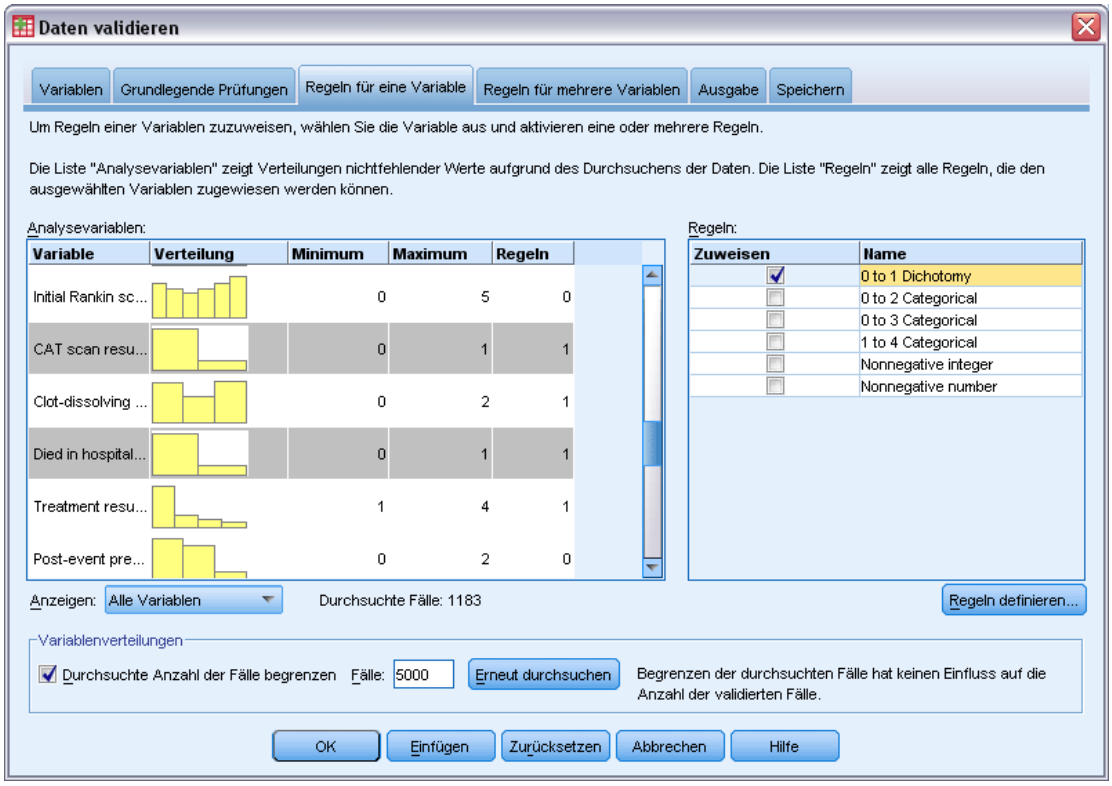
- Um die Daten in *stroke_invalid.sav* auf der Grundlage der kopierten Regeln zu validieren, klicken Sie auf der Symbolleiste auf die Schaltfläche "Zuletzt verwendete Dialogfelder" und wählen Sie Daten validieren aus.

- Klicken Sie auf die Registerkarte Regeln für eine Variable.

In der Liste "Analysevariablen" werden die Variablen, die Sie auf der Registerkarte "Variablen" ausgewählt haben, zusammenfassende Informationen zu deren Verteilungen und die Anzahl der Regeln angezeigt, die ihnen jeweils zugeordnet sind. Variablen, deren Eigenschaften aus der Datei *patient_los.sav* kopiert wurden, besitzen zugeordnete Regeln.

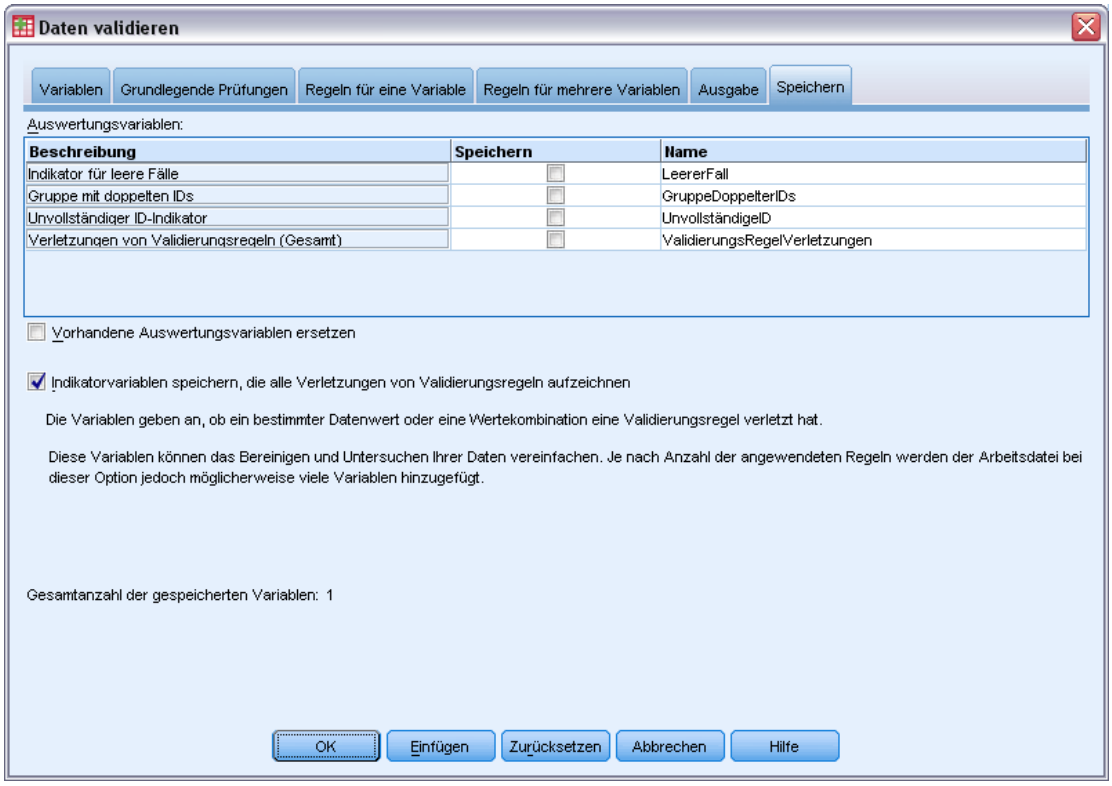
In der Liste "Regeln" werden die Validierungsregeln für eine Variable angezeigt, die in der Datendatei verfügbar sind. Diese Regeln wurden aus der Datei *patient_los.sav* kopiert. Beachten Sie, dass einige dieser Regeln auch auf Variablen zutreffen, für die in der anderen Datendatei keine exakten Entsprechungen vorliegen.

Abbildung 7-11
Dialogfeld "Daten validieren", Registerkarte "Regeln für eine Variable"



- Wählen Sie *Atrial fibrillation*, *History of transient ischemic attack*, *CAT scan result* und *Died in hospital* aus und wenden Sie die Regel 0 to 1 Dichotomy an.
- Wenden Sie 0 to 3 Categorical auf *Post-event rehabilitation* an.
- Wenden Sie 0 to 2 Categorical auf *Post-event preventative surgery* an.
- Wenden Sie Nonnegative integer auf *Length of stay for rehabilitation* an.
- Wenden Sie 1 to 4 Categorical auf die Variablen von *Recoded Barthel index at 1 month* bis *Recoded Barthel index at 6 months* an.
- Klicken Sie auf die Registerkarte Speichern.

Abbildung 7-12
Dialogfeld "Daten validieren", Registerkarte "Speichern"



- Wählen Sie Indikatorvariablen speichern, die alle Verletzungen von Validierungsregeln aufzeichnen aus. Dies vereinfacht es, eine Verbindung zwischen Fällen und Variablen herzustellen, bei denen Validierungsregeln für eine Variable verletzt werden.
- Klicken Sie auf OK.

Regelbeschreibung

Abbildung 7-13
Regelbeschreibung

Regel	Beschreibung
Nonnegative integer	Type: Numeric Domain: Range Flag user-missing values: No Flag system-missing values: Yes Minimum: 0 Flag unlabeled values within range: No Flag noninteger values within range: Yes $VD.SRule[5]: Rule
0 to 1 Dichotomy	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: Yes List: 0; 1 $VD.SRule[1]: Rule
1 to 4 Categorical	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: Yes List: 1; 2; 3; 4 $VD.SRule[4]: Rule

Es werden alle Regeln gezeigt, die mindestens einmal verletzt wurden.

Die Tabelle “Regelbeschreibung” enthält Erklärungen zu den Regeln, die verletzt wurden. Dies ist nützlich, wenn viele Validierungsregeln vorliegen.

Variablenauswertung

Abbildung 7-14
Variablenauswertung

	Regel	Anzahl der Verletzungen
Age category	1 to 4 Categorical	1
	Gesamt	1
Gender	0 to 1 Dichotomy	1
	Gesamt	1
History of angina	0 to 1 Dichotomy	1
	Gesamt	1
Time to hospital	Nonnegative integer	2
	Gesamt	2
Dead on arrival	0 to 1 Dichotomy	1
	Gesamt	1

Die Tabelle “Variablenauswertung” enthält alle Variablen, die mindestens eine Validierungsregel verletzt haben, die verletzten Regeln und die Anzahl der Verletzungen pro Regel und pro Variable.

Fallbericht

Abbildung 7-15
Fallbericht

Fall	Verletzungen von Eine Variable[a]	Identifizierung hospid	patid	physid
175	0 to 1 Dichotomy (1)	OZN	0333204686	883285
274	0 to 1 Dichotomy (1)	OZN	1038840465	103254
310	Nonnegative integer (1)	OZN	2090290204	883285
437	0 to 1 Dichotomy (1)	WPA	2349729006	723384
752	Nonnegative integer (1)	GFG	4993307441	828754
1173	1 to 4 Categorical (1)	ALK	8737661990	185787

a. The number of variables that violated the rule follows each rule.

In der Tabelle "Fallbericht" werden alle Fälle (sowohl nach Fallnummer als auch nach Fallbezeichner), die mindestens eine Validierungsregel verletzt haben, die verletzten Regeln und die Anzahl der Regelverletzungen nach Fall aufgeführt. Die ungültigen Werte werden nun im Daten-Editor angezeigt.

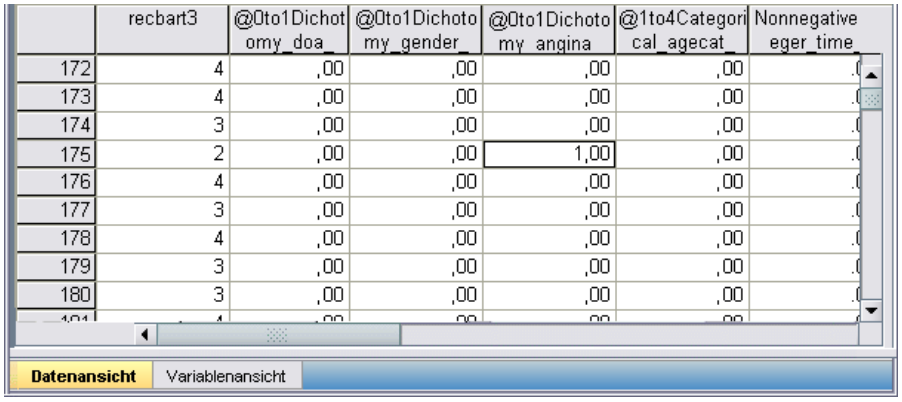
Abbildung 7-16
Daten-Editor mit gespeicherten Indikatorvariablen für Regelverletzungen

	recbart3	@0to3Categorical_clotsolv_	@0to3Categorical_rehab_	@0to1Dichotomy_obesity_	@0to1Dichotomy_dhosp_	@0to1Dichotomy_tia_	@0: oto:
1	4	,00	,00	,00	,00	,00	
2	4	,00	,00	,00	,00	,00	
3	1	,00	,00	,00	,00	,00	
4	4	,00	,00	,00	,00	,00	
5	3	,00	,00	,00	,00	,00	
6	4	,00	,00	,00	,00	,00	
7	4	,00	,00	,00	,00	,00	
8	4	,00	,00	,00	,00	,00	
9	4	,00	,00	,00	,00	,00	

Datenansicht | Variablenansicht

Für jede Anwendung einer Validierungsregel wird eine separate Indikatorvariable erstellt. So entstpricht *@0to3Categorical_clotsolv_* der Anwendung der Validierungsregel "0 to 3 Categorical" auf die Variable *Clot-dissolving drugs*. Wenn Sie bei einem Fall feststellen möchten, welche Variable einen ungültigen Wert aufweist, betrachten Sie am besten die Werte der Indikatorvariablen. Der Wert 1 bedeutet, dass der Wert der zugeordneten Variablen ungültig ist.

Abbildung 7-17
Daten-Editor mit Indikatorvariable für Regelverletzung in Fall 175

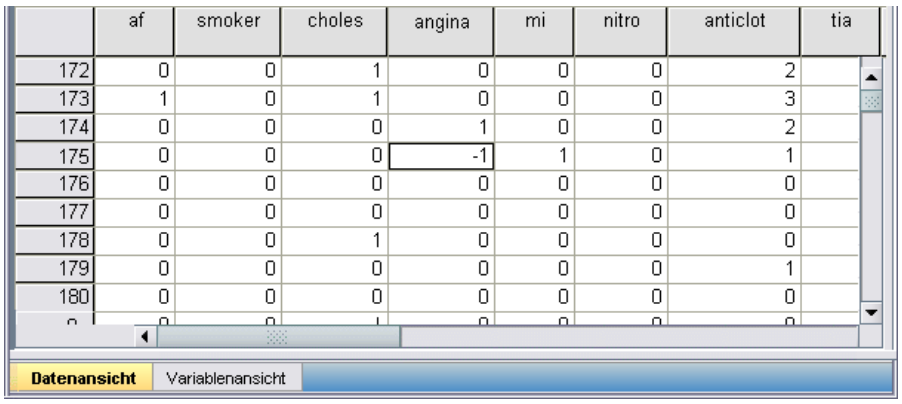


	recbart3	@0to1Dichotomy_doa_	@0to1Dichotomy_gender_	@0to1Dichotomy_angina	@1to4Categorical_agecat_	Nonnegativeeger_time_
172	4	,00	,00	,00	,00	
173	4	,00	,00	,00	,00	
174	3	,00	,00	,00	,00	
175	2	,00	,00	1,00	,00	
176	4	,00	,00	,00	,00	
177	3	,00	,00	,00	,00	
178	4	,00	,00	,00	,00	
179	3	,00	,00	,00	,00	
180	3	,00	,00	,00	,00	

Datenansicht | Variablenansicht

Wechseln Sie zu Fall 175, dem ersten Fall, bei dem eine Regelverletzung auftritt. Um die Suche zu beschleunigen, betrachten Sie die Indikatorvariablen, die den Variablen in der Tabelle "Variablenauswertung" zugeordnet sind. Es ist offensichtlich, dass *History of angina* einen ungültigen Wert aufweist.

Abbildung 7-18
Daten-Editor mit ungültigem Wert für "History of angina"



	af	smoker	choles	angina	mi	nitro	anticlot	tia
172	0	0	1	0	0	0	2	
173	1	0	1	0	0	0	3	
174	0	0	0	1	0	0	2	
175	0	0	0	-1	1	0	1	
176	0	0	0	0	0	0	0	
177	0	0	0	0	0	0	0	
178	0	0	1	0	0	0	0	
179	0	0	0	0	0	0	1	
180	0	0	0	0	0	0	0	

Datenansicht | Variablenansicht

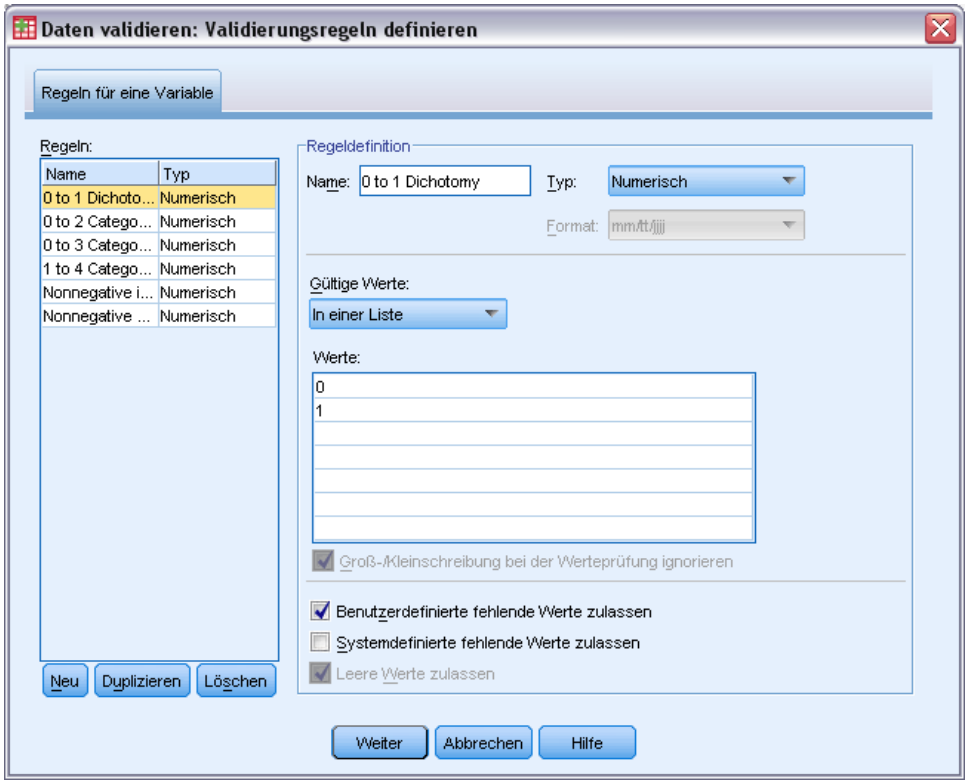
History of angina weist den Wert –1 auf. Dieser Wert ist zwar ein gültiger fehlender Wert für die Behandlungs- und Ergebnisvariablen in der Datendatei, an der vorliegenden Stelle ist er jedoch ungültig, weil für die Anamnesevariablen keine benutzerdefiniert fehlenden Werte festgelegt wurden.

Definieren von eigenen Regeln

Die aus der Datei *patient_los.sav* kopierten Validierungsregeln sind zweifellos nützlich, reichen jedoch nicht aus. Es gibt Situationen, in denen Patienten, die bereits vor der Ankunft verstorben waren, versehentlich als im Krankenhaus verstorben erfasst werden. Eine Situation dieser Art kann nicht mit einer Regel für eine Variable erkannt werden; Sie benötigen eine Regel für mehrere Variablen.

- Klicken Sie auf der Symbolleiste auf das Symbol “Zuletzt verwendete Dialogfelder” und wählen Sie Daten validieren aus.
- Klicken Sie auf die Registerkarte Regeln für eine Variable. (Sie müssen Regeln für *Hospital size*, die Variablen für die Rankin-Scores und die Variablen der nicht umkodierten Barthel-Indizes erstellen.)
- Klicken Sie auf Regeln definieren.

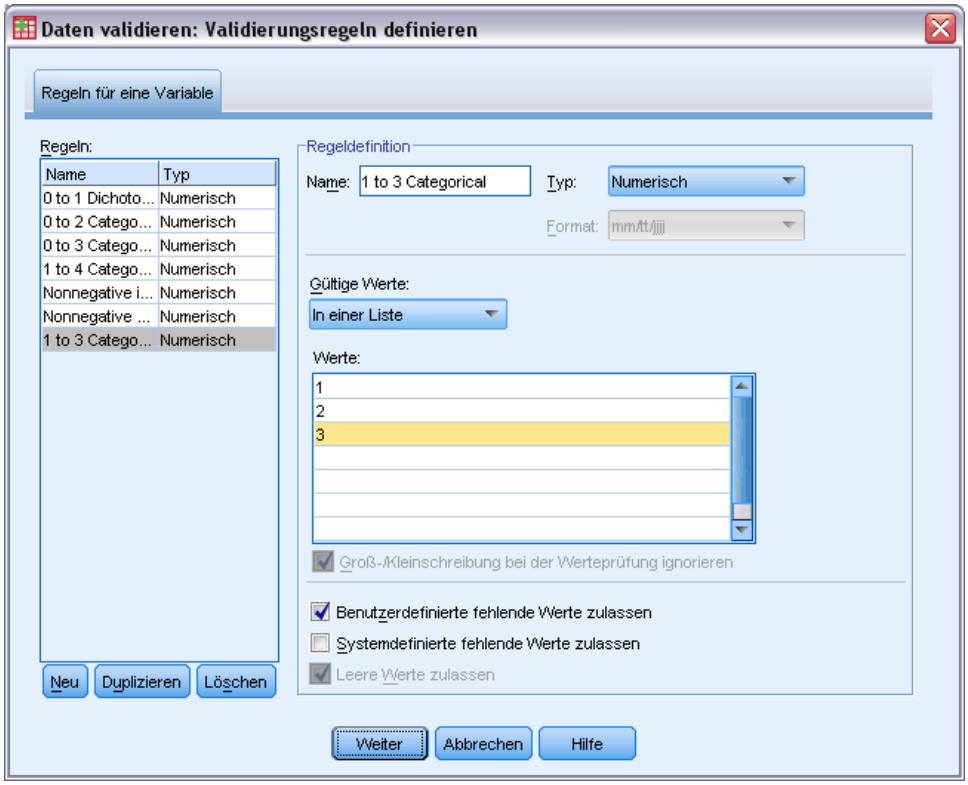
Abbildung 7-19
Dialogfeld “Validierungsregeln definieren”, Registerkarte “Regeln für eine Variable”



In der Liste “Regeln” werden die aktuell definierten Regeln angezeigt. Die Regel 0 to 1 Dichotomy ist ausgewählt, und ihre Eigenschaften werden im Gruppenfeld “Regeldefinition” angezeigt.

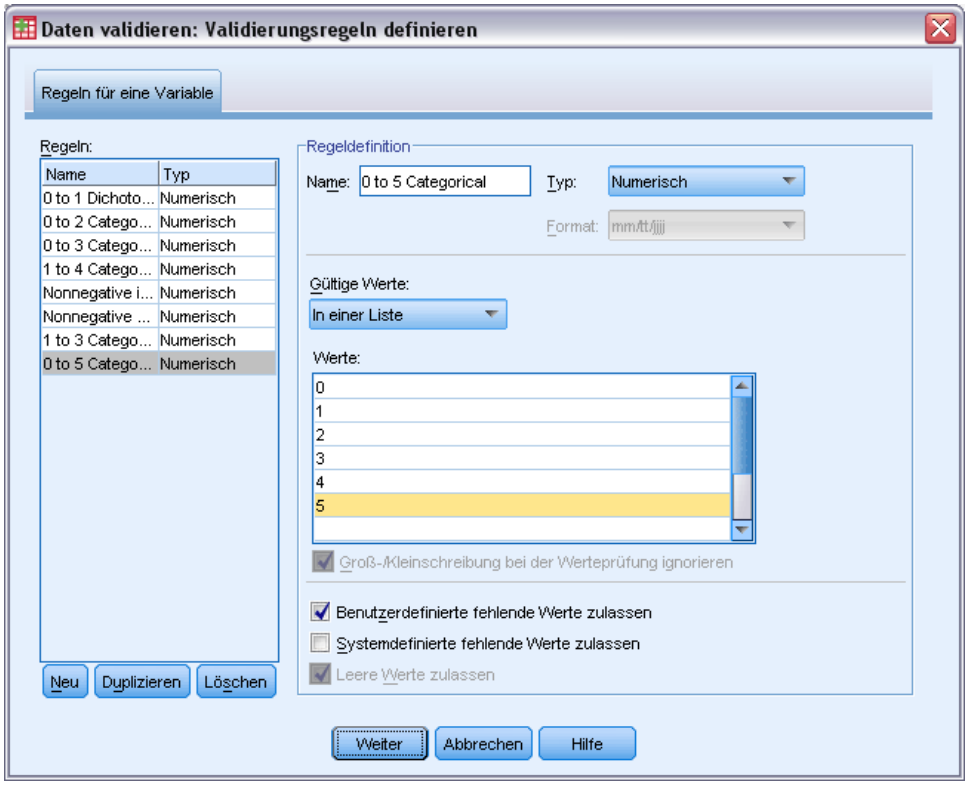
- Um eine Regel zu definieren, klicken Sie auf Neu.

Abbildung 7-20
Dialogfeld "Validierungsregeln definieren", Registerkarte "Regeln für eine Variable" (Definition von "1 to 3 Categorical")



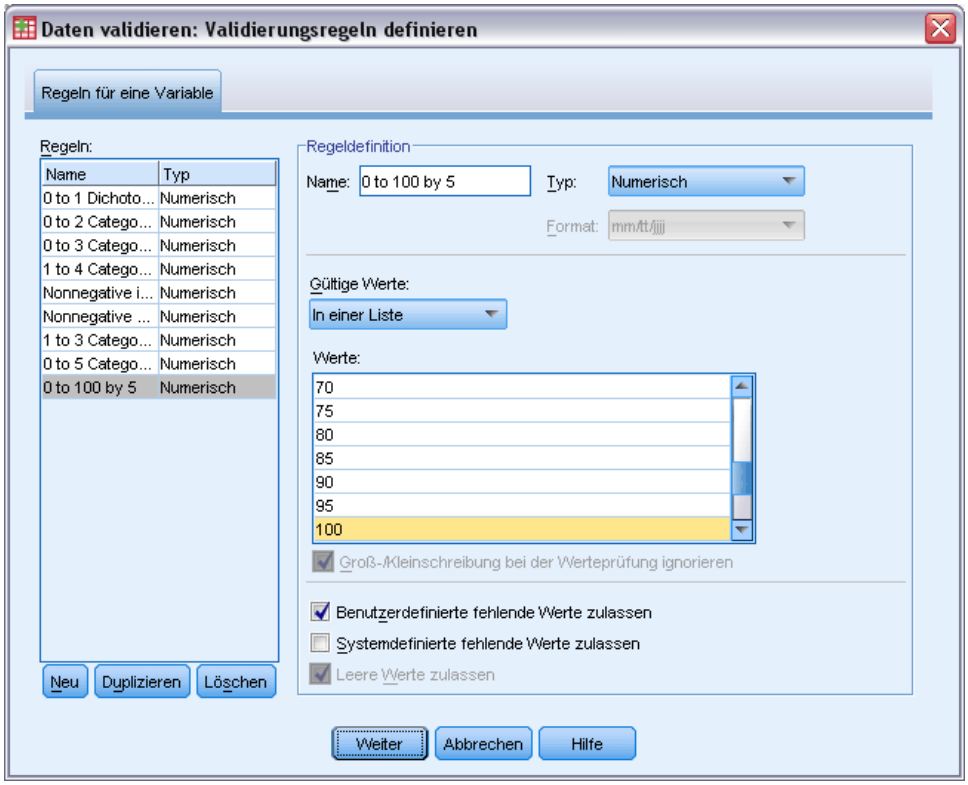
- Geben Sie als Name der Regel 1 to 3 Categorical ein.
- Wählen Sie im Feld "Gültige Werte" den Eintrag In einer Liste aus.
- Geben Sie die Werte 1, 2 und 3 ein.
- Deaktivieren Sie Systemdefinierte fehlende Werte zulassen.
- Um die Regel für die Rankin-Scores zu definieren, klicken Sie auf Neu.

Abbildung 7-21
Dialogfeld "Validierungsregeln definieren", Registerkarte "Regeln für eine Variable" (Definition von "0 to 5 Categorical")



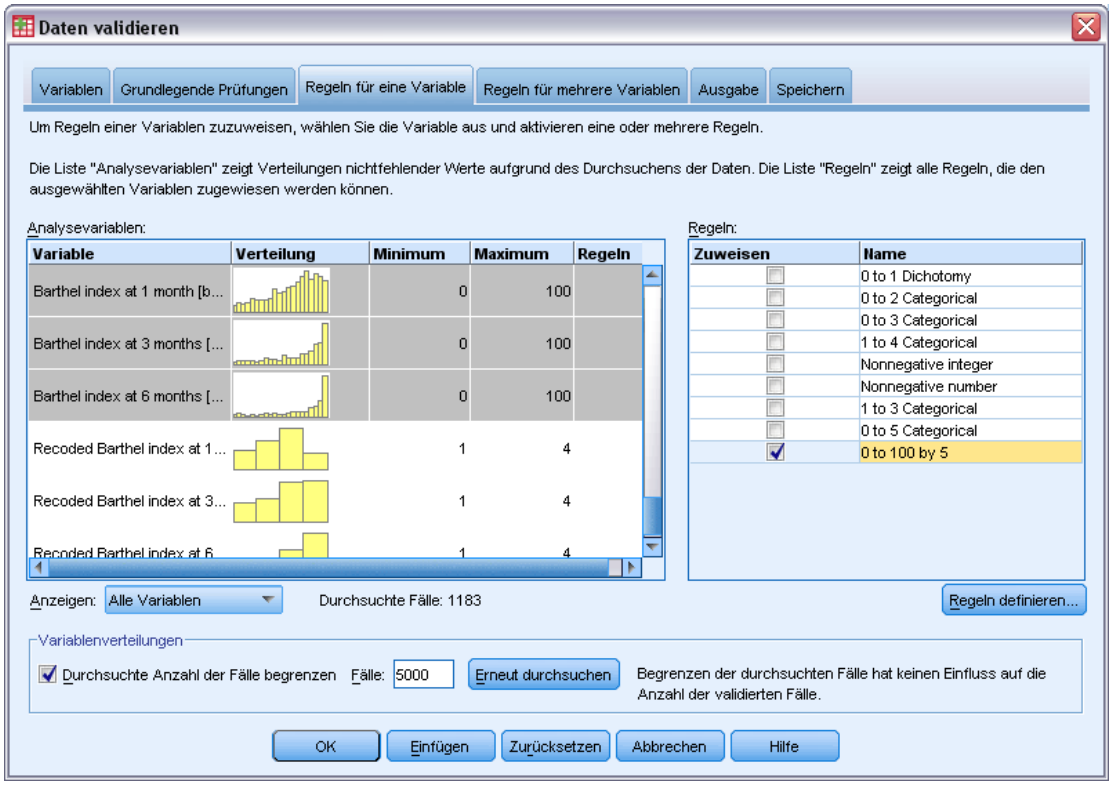
- Geben Sie als Name der Regel 0 to 5 Categorical ein.
- Wählen Sie im Feld "Gültige Werte" den Eintrag In einer Liste aus.
- Geben Sie die Werte 0, 1, 2, 3, 4 und 5 ein.
- Deaktivieren Sie Systemdefinierte fehlende Werte zulassen.
- Um die Regel für die Barthel-Indizes zu definieren, klicken Sie auf Neu.

Abbildung 7-22
Dialogfeld "Validierungsregeln definieren", Registerkarte "Regeln für eine Variable" (Definition von "0 to 100 by 5")



- Geben Sie als Name der Regel 0 to 100 by 5 ein.
- Wählen Sie im Feld "Gültige Werte" den Eintrag In einer Liste aus.
- Geben Sie die Werte 0, 5, ... bis 100 ein.
- Deaktivieren Sie Systemdefinierte fehlende Werte zulassen.
- Klicken Sie auf Weiter.

Abbildung 7-23
Dialogfeld "Daten validieren", Registerkarte "Regeln für eine Variable" (Definition von "0 to 100 by 5")



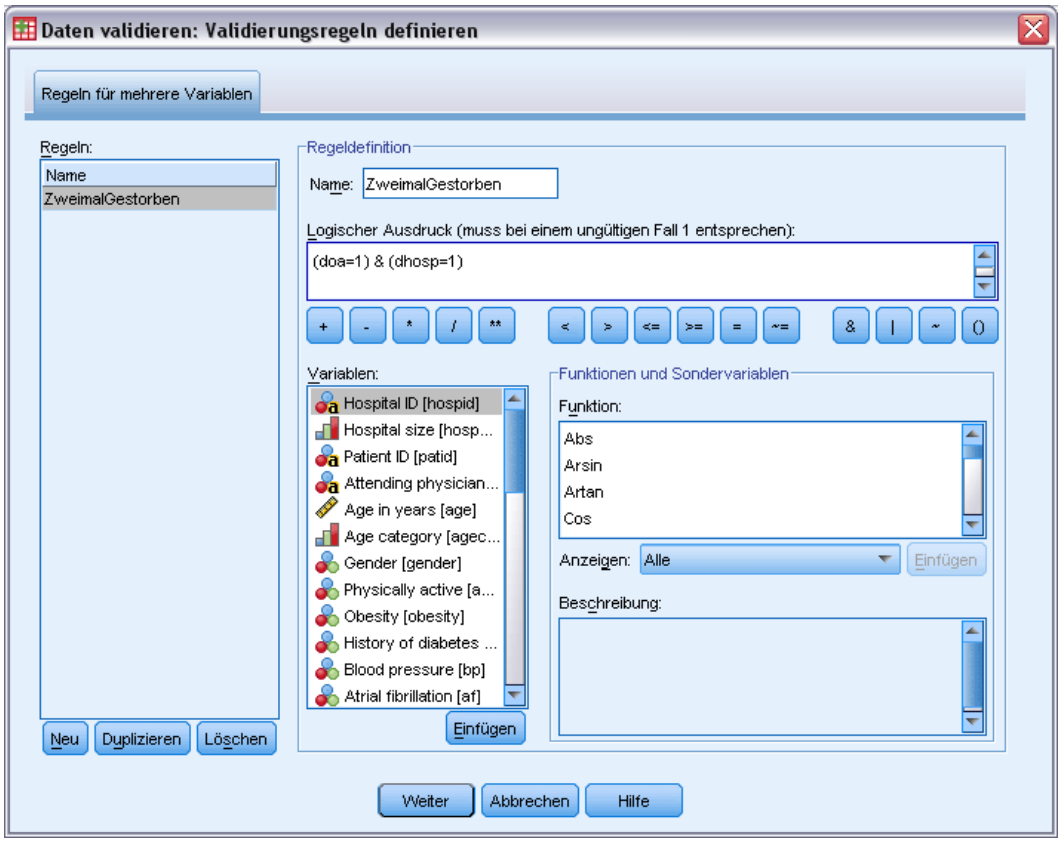
Jetzt müssen Sie die definierten Regeln Variablen zuordnen.

- Wenden Sie 1 to 3 Categorical auf *Hospital size* an.
- Wenden Sie 0 to 5 Categorical auf *Initial Rankin score* sowie die Variablen von *Rankin score at 1 month* bis *Rankin score at 6 months* an.
- Wenden Sie 0 to 100 by 5 auf die Variablen von *Barthel index at 1 month* bis *Barthel index at 6 months* an.
- Klicken Sie auf die Registerkarte Regeln für mehrere Variablen.

Gegenwärtig sind keine Regeln definiert.

- Klicken Sie auf Regeln definieren.

Abbildung 7-24
Dialogfeld "Validierungsregeln definieren", Registerkarte "Regeln für mehrere Variablen"



Wenn keine Regeln vorliegen, wird automatisch eine neue Platzhalterrregel erstellt.

- Geben Sie als Name der Regel ZweimalGestorben ein.
- Geben Sie als logischen Ausdruck (doa=1) & (dhosp=1) ein. Dieser Ausdruck ergibt den Wert 1, wenn für den Patienten sowohl der Tod vor der Ankunft als auch der Tod im Krankenhaus aufgezeichnet wurde.
- Klicken Sie auf Weiter.

  Die neue Regel auf der Registerkarte "Regeln für mehrere Variablen" wird automatisch ausgewählt.
- Klicken Sie auf OK.

Regeln für mehrere Variablen

Abbildung 7-25
Regeln für mehrere Variablen

Regel	Anzahl der Verletzungen	Ausdruck
Zweimal Gestorben	27	(doa = 1) & (dhosp = 1)

Die Liste der Regeln für mehrere Variablen enthält Regeln, die mindestens einmal verletzt wurden, die Anzahl der Verletzungen und eine Beschreibung jeder verletzten Regel.

Fallbericht

Abbildung 7-26
Fallbericht

	Validation Rule Violations		Identifizierung		
Fall	Single-Variable[a]	Cross-Variable	hospid	patid	physid
20		Zweimal Gestorben	PBW	1192970826	355184
49		Zweimal Gestorben	NHV	8717862852	237418
129		Zweimal Gestorben	QWS	6901932085	215041
138		Zweimal Gestorben	RLD	1205005069	695521
162		Zweimal Gestorben	OZN	5546809538	125304
175	0 to 1 Dichotomy (1)		OZN	0333204686	883285
274	0 to 1 Dichotomy (1)		OZN	1038840465	103254
310	Nonnegative integer (1)		OZN	2090290204	883285
414		Zweimal Gestorben	WPA	3351107142	462020
437	0 to 1 Dichotomy (1)		WPA	2349729006	723384
447		Zweimal Gestorben	WPA	7163481282	519548
458		Zweimal Gestorben	WPA	9159094175	652070
462		Zweimal Gestorben	WPA	2137520354	723384
537		Zweimal Gestorben	SLB	5246122506	928076
544		Zweimal Gestorben	SLB	1605957462	506108
620		Zweimal Gestorben	GFG	8141858966	828754
629		Zweimal Gestorben	GFG	3397891610	539412
630		Zweimal Gestorben	GFG	3397891610	539412
639		Zweimal Gestorben	GFG	3962622031	327422
644		Zweimal Gestorben	GFG	4271782383	749432
649		Zweimal Gestorben	GFG	0950686750	618069
653		Zweimal Gestorben	GFG	0663642766	001448
722		Zweimal Gestorben	GFG	0418125590	877354
748		Zweimal Gestorben	GFG	8744721380	539412
752	Nonnegative integer (1) 0 to 1 Dichotomy (3)		GFG	4993307441	828754
868		Zweimal Gestorben	WWL	9714672452	237547
881		Zweimal Gestorben	WWL	6613279456	574275
915		Zweimal Gestorben	EFX	2575793702	501318
933		Zweimal Gestorben	IZO	2807437472	680253
1010		Zweimal Gestorben	BLA	5284009939	657638
1028		Zweimal Gestorben	BLA	8021997463	185703
1054		Zweimal Gestorben	ALK	0950897644	267830
1173	1 to 4 Categorical (1)		ALK	8737661990	185787

a. The number of variables that violated the rule follows each rule.

Der Fallbericht enthält jetzt neben den bereits vorher erkannten Fällen, die die Regeln für eine Variable verletzen, auch die Fälle, die die Regeln für mehrere Variablen verletzen. Diese Fälle müssen den für die Datenerfassung zuständigen Personen gemeldet werden, damit sie korrigiert werden können.

Zusammenfassung

Die Analytikerin verfügt jetzt über die Informationen für einen vorläufigen Bericht an den Leiter der Datenerfassung.

Verwandte Prozeduren

Die Prozedur “Daten validieren” ist nützlich für die Qualitätskontrolle der Daten.

- Mit der Prozedur Ungewöhnliche Fälle identifizieren können Sie Muster in den Daten analysieren und Fälle identifizieren, bei denen einige signifikante Werte abweichen.

Kapitel
8

Automatisierte Datenaufbereitung

Die Aufbereitung von Daten zur Analyse ist einer der wichtigsten Schritte in jedem Projekt – und gewöhnlich auch einer der zeitaufwendigsten. Die automatisierte Datenaufbereitung (ADP) übernimmt diese Aufgabe für Sie. Sie analysiert Ihre Daten und identifiziert Problemlösungen, findet problematische oder wahrscheinlich nicht nützliche Felder, leitet zum passenden Zeitpunkt neue Attribute ab und verbessert die Leistungsfähigkeit durch intelligente Screening-Methoden. Sie können den Algorithmus **vollautomatisch** verwenden und so Problemlösungen auswählen und anwenden oder Sie können ihn **interaktiv** verwenden und so die Änderungen in einer Vorschau betrachten, bevor sie vorgenommen werden, und sie gegebenenfalls akzeptieren oder ablehnen.

Mit ADP können Sie Ihre Daten schnell und einfach für die Modellerstellung aufbereiten, ohne über Vorkenntnisse der dazugehörigen statistischen Konzepte verfügen zu müssen. Modelle lassen sich damit schneller erstellen und scoren; zudem verbessert sich mit ADP die Robustheit automatisierter Modellierungsprozesse.

Interaktive Verwendung der automatisierten Datenaufbereitung

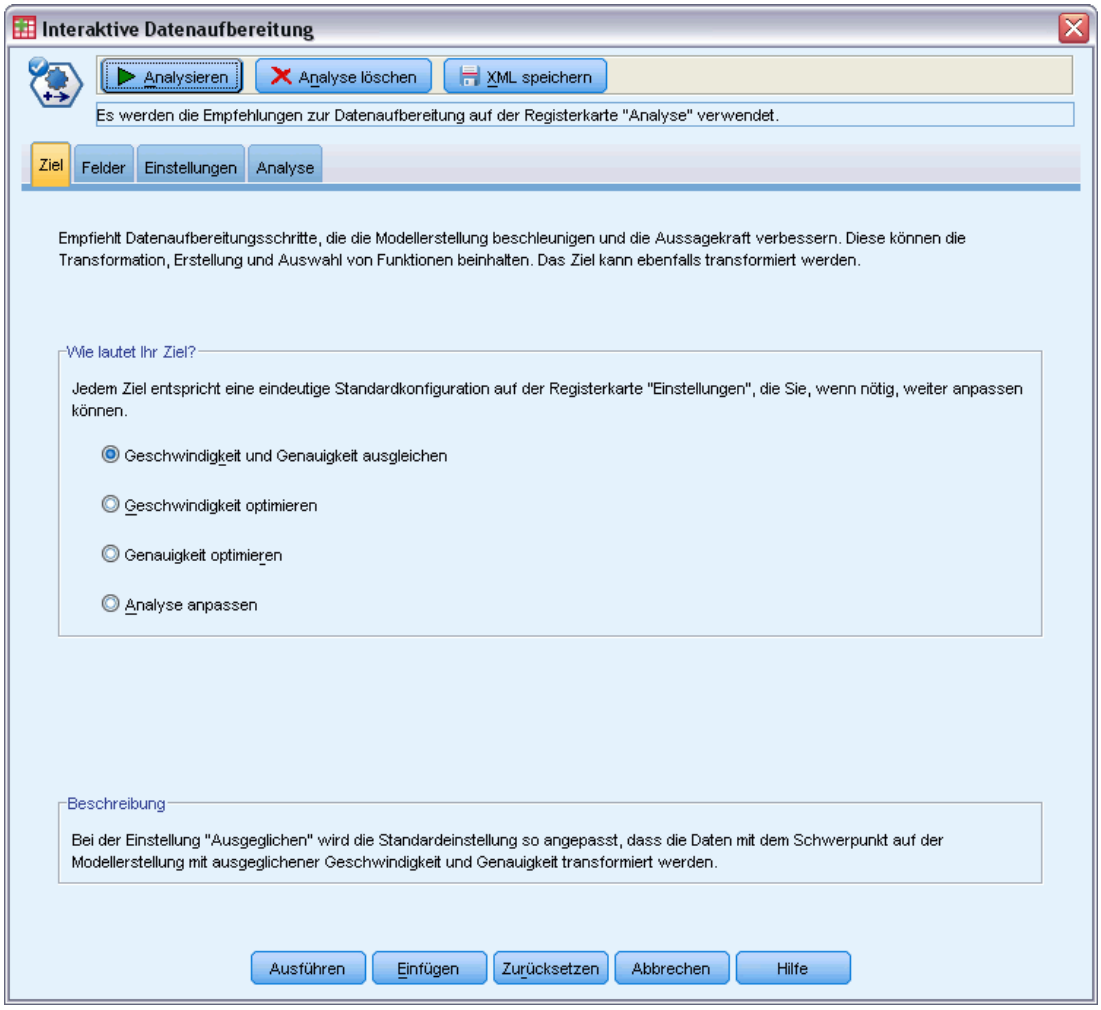
Eine Versicherungsgesellschaft mit beschränkten Ressourcen für die Untersuchung der Versicherungsansprüche von Hauseigentümern möchte ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen. Sie haben ein Datenbeispiel früherer Ansprüche unter *insurance_claims.sav* zusammengestellt. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 139. Vor Erstellung des Modells bereiten sie die Daten für die Modellierung mithilfe der automatisierten Datenaufbereitung vor. Da sie die vorgeschlagenen Transformationen zunächst überprüfen möchten, bevor die Transformationen angewendet werden, nutzen sie die automatisierte Datenaufbereitung im interaktiven Modus.

Auswahl aus Objekten

- Zur interaktiven Ausführung der automatisierten Datenaufbereitung wählen Sie aus den Menüs:

Transformieren
  Daten für Modellierung vorbereiten
    Interaktiv...

Abbildung 8-1
Registerkarte Ziel


Interaktive Datenaufbereitung
Analysieren
Analyse löschen
XML speichern
Es werden die Empfehlungen zur Datenaufbereitung auf der Registerkarte "Analyse" verwendet.
Ziel
Felder
Einstellungen
Analyse
Empfiehlt Datenaufbereitungsschritte, die die Modellerstellung beschleunigen und die Aussagekraft verbessern. Diese können die Transformation, Erstellung und Auswahl von Funktionen beinhalten. Das Ziel kann ebenfalls transformiert werden.
Wie lautet Ihr Ziel?
Jedem Ziel entspricht eine eindeutige Standardkonfiguration auf der Registerkarte "Einstellungen", die Sie, wenn nötig, weiter anpassen können.
Geschwindigkeit und Genauigkeit ausgleichen
Geschwindigkeit optimieren
Genauigkeit optimieren
Analyse anpassen
Beschreibung
Bei der Einstellung "Ausgeglichen" wird die Standardeinstellung so angepasst, dass die Daten mit dem Schwerpunkt auf der Modellerstellung mit ausgeglichener Geschwindigkeit und Genauigkeit transformiert werden.
Ausführen
Einfügen
Zurücksetzen
Abbrechen
Hilfe

Die erste Registerkarte fragt nach einem Ziel, das die Standardeinstellungen regelt. Doch was ist der faktische Unterschied zwischen den Zielen? Wir führen die Prozedur mit jedem einzelnen Ziel durch und sehen, wie sich die Ergebnisse unterscheiden.

- Stellen Sie sicher, dass Geschwindigkeit & Genauigkeit ausgleichen ausgewählt ist, und klicken Sie auf Analysieren.

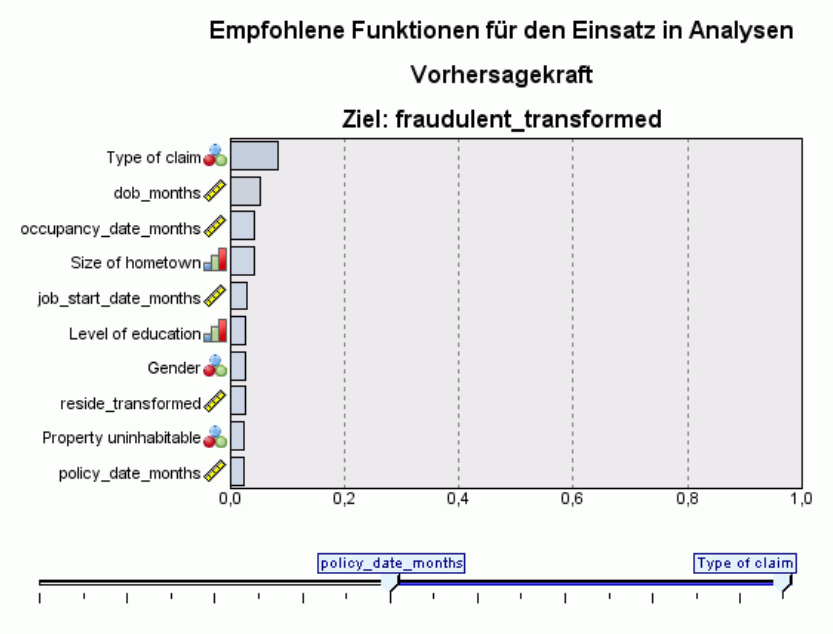
Abbildung 8-2
Registerkarte "Analyse", Feldverarbeitungsübersicht für ausgeglichene Ziele

Feldverarbeitungsübersicht

Felder		X
Ziel		1
Eingabefunktionen		18
Empfohlene Funktionen für den Einsatz in Analysen	Gesamtergebnis	18
	Originalfelder (nicht transformiert)	8
	Transformationen von Originalfeldern	5
	Aus Datums- und Zeitangaben abgeleitet	5
	Erstellt	0
Nicht verwendete Eingabefunktionen		0

Während die Daten verarbeitet werden, richtet sich die Konzentration automatisch auf die Registerkarte "Analyse". Die Standardhauptansicht ist die Feldverarbeitungsübersicht, die einen Überblick darüber gibt, wie die Felder von der automatisierten Datenaufbereitung verarbeitet wurden. Es gibt ein Einzelziel, 18 Eingaben und 18 für die Modellerstellung empfohlene Felder. Von den für die Modellierung empfohlenen Feldern sind acht originale Eingabefelder, fünf sind Transformationen originaler Eingabefelder und fünf sind von Datum- und Uhrzeitfeldern abgeleitet.

Abbildung 8-3
Registerkarte "Analyse," Vorhersagekraft bei "ausgeglichenen Zielen"



Als Hilfsansicht wird standardmäßig die Vorhersagekraft angezeigt, die einen schnellen Überblick darüber gibt, welche empfohlenen Felder für die Modellerstellung am nützlichsten sind. Hinweis: Zwar werden 18 Einflussgrößen für die Analyse empfohlen, doch werden standardmäßig nur die ersten zehn im Vorhersagekraft-Diagramm angezeigt. Mehr oder weniger Felder können mit dem Schieberegler unterhalb der Grafik angezeigt werden.

Wird Geschwindigkeit & Genauigkeit ausgleichen als Ziel ausgewählt, dann wird *Anspruchstyp* als "beste" Einflussgröße identifiziert, gefolgt vom aktuellen Alter des Anspruchsberechtigten in Monaten (die berechnete Zeitspanne vom Geburtsdatum bis zum aktuellen Datum) und die Anzahl der Monate seit dem Zeitpunkt, zu dem der Anspruchsberechtigte in den Besitz des Eigentums gelangt ist (die berechnete Zeitspanne vom Datum der Besitznahme bis zum aktuellen Datum).

- Klicken Sie auf Analyse löschen und anschließend auf die Registerkarte "Ziele".
- Wählen Sie Geschwindigkeit optimieren und klicken Sie auf Analysieren.

Abbildung 8-4
Registerkarte "Analyse", Feldbearbeitungsübersicht bei "optimierter Geschwindigkeit"

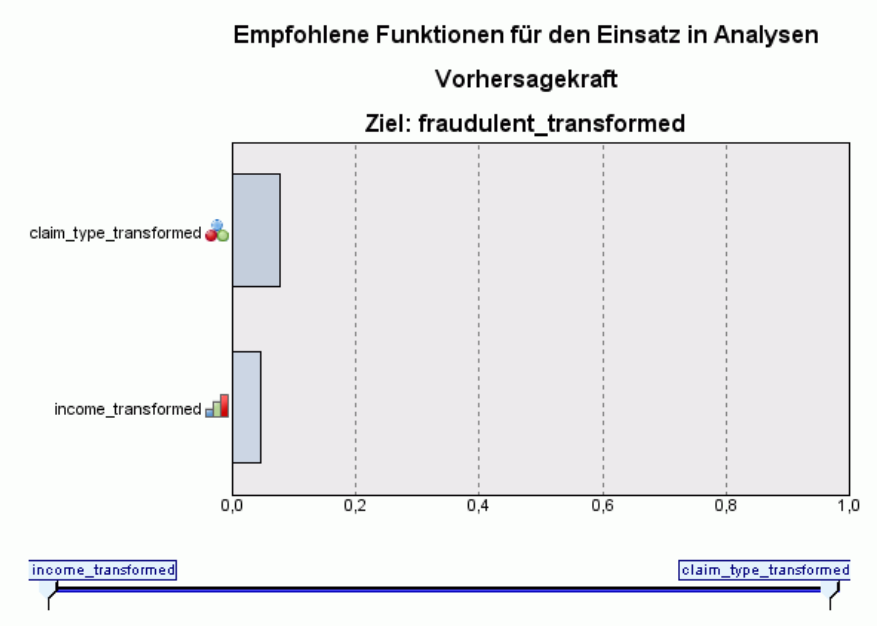
Feldverarbeitungsübersicht

Felder		X
Ziel		1
Eingabefunktionen		18
Empfohlene Funktionen für den Einsatz in Analysen	Gesamtergebnis	2
	Originalfelder (nicht transformiert)	0
	Transformationen von Originalfeldern	2
	Aus Datums- und Zeitangaben abgeleitet	0
	Erstellt	0
Nicht verwendete Eingabefunktionen		16

- Es wurde eine Funktionserstellung angefordert, aber es konnten keine Vorhersagefunktionen erstellt werden. Die häufigsten Gründe sind: zu wenige stetige Eingabefunktionen mit hohem Zielzusammenhang oder alle stetigen Eingabefunktionen waren unabhängig.

Während die Daten verarbeitet werden, richtet sich die Konzentration automatisch wieder auf die Registerkarte "Analyse". In diesem Fall werden nur zwei Felder für die Modellerstellung empfohlen und beide sind Transformationen der originalen Felder.

Abbildung 8-5
Registerkarte "Analyse," Vorhersagekraft bei "optimierter Geschwindigkeit"



Wird Geschwindigkeit optimieren als Ziel eingegeben, dann wird *claim_type_transformed* gefolgt von *income_transformed* als "beste" Einflussgröße identifiziert.

- Klicken Sie auf Analyse löschen und anschließend auf die Registerkarte "Ziele".
- Wählen Sie Genauigkeit optimieren und klicken Sie auf Analysieren.

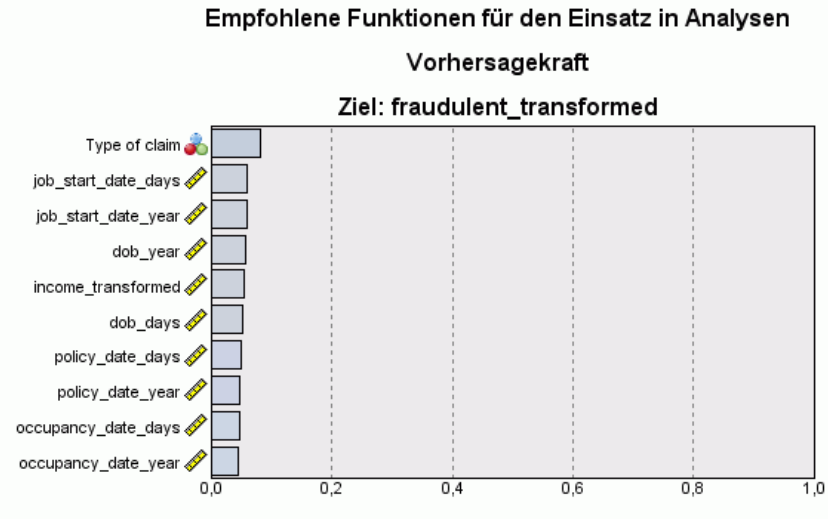
Abbildung 8-6
Registerkarte "Analyse", Vorhersagekraft bei "optimierter Genauigkeit"

Feldverarbeitungsübersicht

Felder		X
Ziel		1
Eingabefunktionen		18
Empfohlene Funktionen für den Einsatz in Analysen	Gesamtergebnis	32
	Originalfelder (nicht transformiert)	8
	Transformationen von Originalfeldern	5
	Aus Datums- und Zeitangaben abgeleitet	19
	Erstellt	0
Nicht verwendete Eingabefunktionen		0

Wird Genauigkeit optimieren als Ziel eingegeben, werden 32 Felder für die Modellerstellung empfohlen, da mehr Felder aus Datumsangaben und Uhrzeiten durch das Extrahieren von Tagen, Monaten und Jahren aus Datumsangaben und Stunden, Minuten und Sekunden aus Uhrzeiten abgeleitet werden.

Abbildung 8-7
Registerkarte "Analyse", Vorhersagekraft bei "optimierter Genauigkeit"



Type of claim (Anspruchstyp) wird als die "beste" Einflussgröße identifiziert, gefolgt von der Anzahl der Tage seit dem letzten Beschäftigungsbeginn des Anspruchsnehmers (die berechnete Zeitspanne seit dem Datum des Beschäftigungsbeginns bis zum aktuellen Datum) und dem Jahr,

in dem der Anspruchsnehmer die aktuelle Beschäftigung aufgenommen hat (extrahiert aus dem Datum des Beschäftigungsbeginns).

Zusammenfassung:

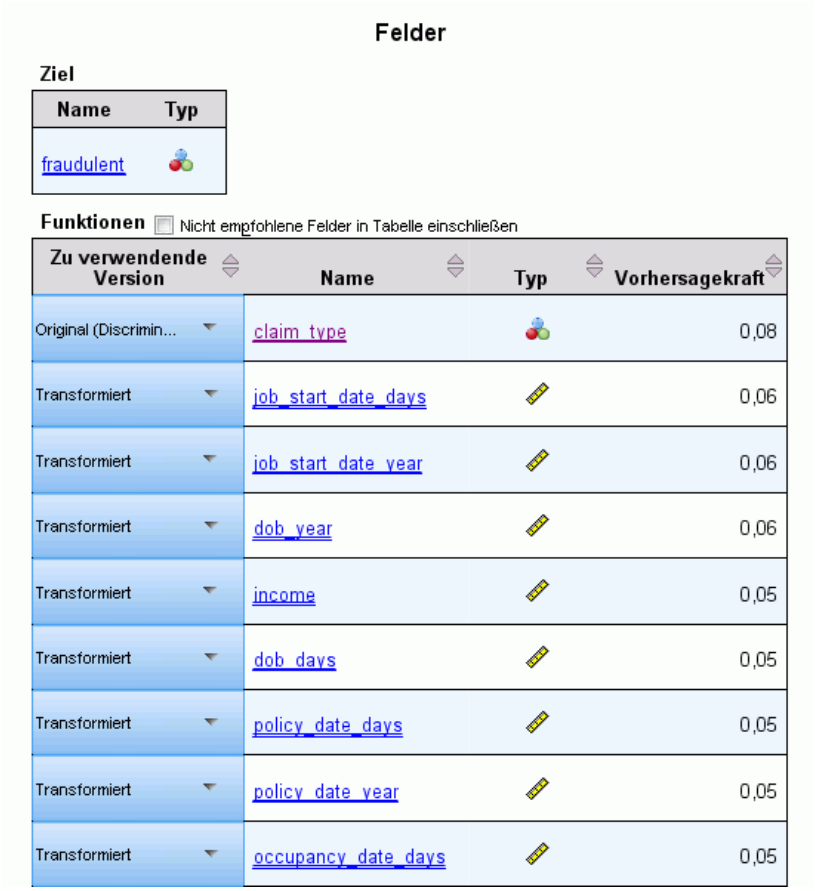
- Geschwindigkeit & Genauigkeit ausgleichen erzeugt für die Modellierung verwendbare Felder aus Daten und transformiert ggf. stetige Felder wie *reside* (Haushaltsgröße), um sie normaler zu verteilen.
- Genauigkeit optimieren erzeugt einige zusätzliche Felder aus Datumsangaben (außerdem werden Ausreißer überprüft und ggf. stetige Ziele für eine normalere Verteilung transformiert).
- Bei Geschwindigkeit optimieren werden keine Datumsangaben aufbereitet und keine stetigen Felder neu skaliert, sondern Kategorien aus kategorialen Einflussgrößen zusammengeführt und stetige Einflussgrößen klassiert, wenn das Ziel kategorial ist (und eine Merkmalsauswahl und -erstellung durchgeführt, wenn das Ziel stetig ist).

Die Versicherungsgesellschaft beschließt, die Ergebnisse bei Genauigkeit optimieren näher zu untersuchen.

▶ Wählen Sie aus der Dropdown-Liste in der Hauptansicht die Option Felder.

Felder und Felddetails

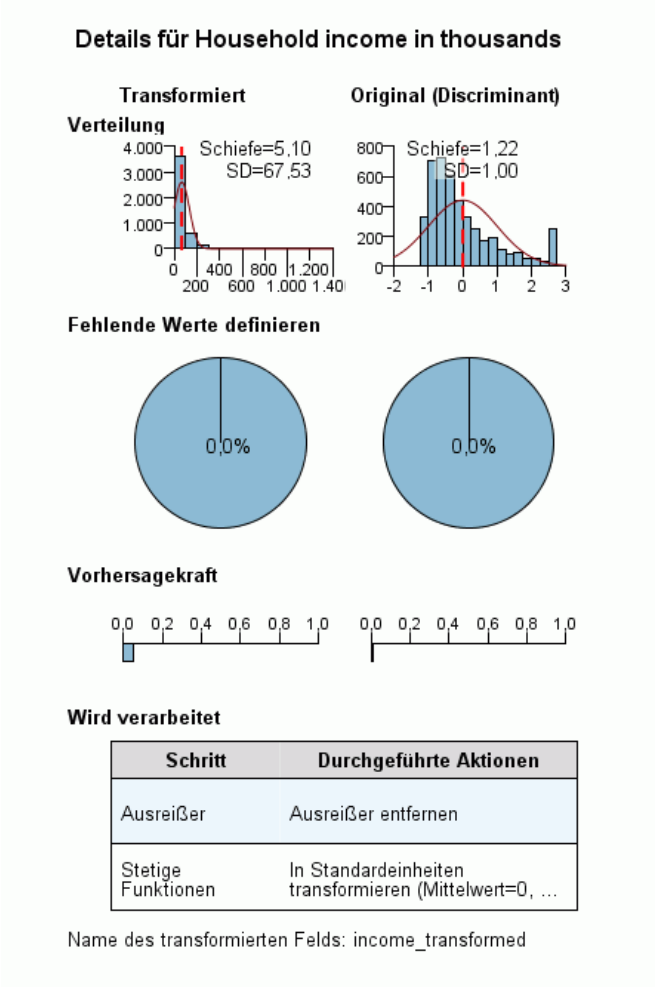
Abbildung 8-8
Felder



In der Ansicht “Felder” werden die verarbeiteten Felder angezeigt sowie ob die ADP diese zur Verwendung bei der Modellerstellung empfiehlt. Durch Klicken auf einen Feldnamen werden in der verknüpften Ansicht weitere Informationen über das Feld angezeigt.

- Klicken Sie auf income (Einkommen).

Abbildung 8-9
Felddetails für "Household income in thousands" (Haushaltseinkommen in Tausend)

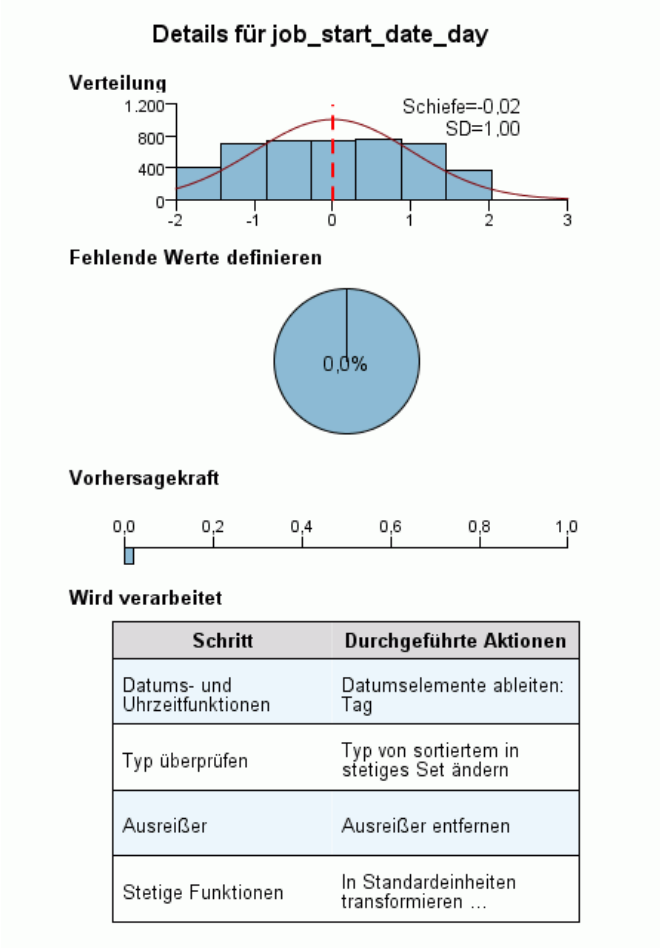


Die Ansicht "Felddetails" zeigt die Verteilung von *Household income in thousands* im Original und in der Transformation. Der Verarbeitungstabelle zufolge wurden Datensätze, die als Ausreißer identifiziert wurden, getrimmt (durch Setzen ihrer Werte gleich dem Trennwert zur Bestimmung von Ausreißern) und das Feld wurde auf einen Mittelwert von 0 und eine Standardabweichung von 1 standardisiert. Die "Beule" am rechten Rand des Histogramms des transformierten Felds zeigt, dass eine Reihe von Datensätzen, vielleicht über 200, als Ausreißer identifiziert wurden. Das Einkommen hat eine sehr schiefe Verteilung, was der Fall sein kann, wenn der Standardtrennwert bei der Bestimmung von Ausreißern zu aggressiv ist.

Auffällig ist auch der Anstieg in der Vorhersagekraft des transformierten Felds gegenüber dem originalen Feld. Es scheint sich um eine nützliche Transformation zu handeln.

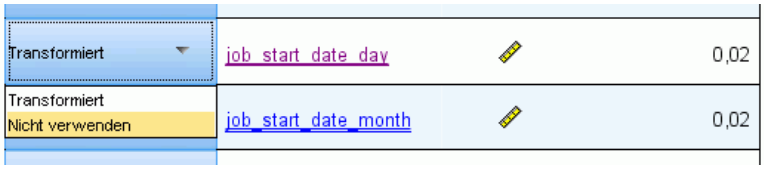
- Klicken Sie in der Ansicht "Felder" auf job_start_date_day (Tag des Arbeitsbeginns). (Hinweis: Nicht zu verwechseln mit *job_start_date_days* (Tage seit Arbeitsbeginn).)

Abbildung 8-10
Felddetails für job_start_date_day



Das Feld *job_start_date_day* ist der extrahierte Tag aus *Employment starting date [job_start_date]* (Beschäftigungsbeginn). Es ist hochgradig unwahrscheinlich, dass dieses Feld eine tatsächliche Aussagekraft darüber hat, ob ein Anspruch betrügerisch ist, und daher möchte es die Versicherungsgesellschaft nicht in die Modellerstellung einbeziehen.

Abbildung 8-11
Felddetails für "Household income in thousands" (Haushaltseinkommen in Tausend)

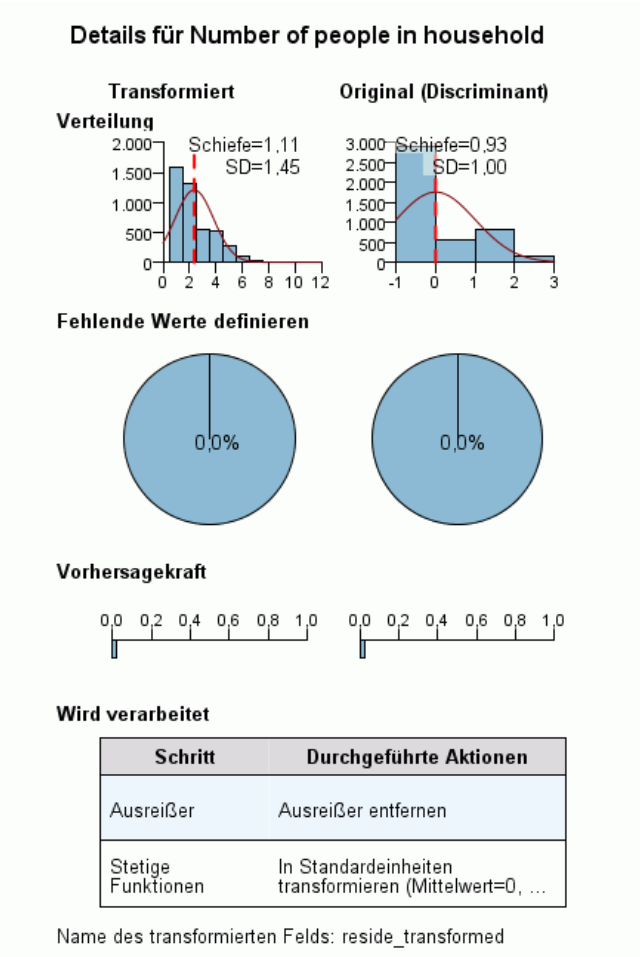


- Wählen Sie in der Ansicht "Felder" Nicht verwenden aus der Dropdown-Liste "Zu verwendende Version" in der Zeile job_start_date_day.

  Führen Sie diesen Vorgang bei allen Feldern mit dem Suffix *_day* und *_month* durch.

- Klicken Sie auf reside in der Ansicht "Felder".

Abbildung 8-12
Felddetails für "Household income in thousands" (Haushaltseinkommen in Tausend)



Die Felddetailsanzeige *Number of people in household* (Haushaltsgröße) zeigt ein interessantes und etwas ungewöhnliches Ergebnis. Den Vorhersagekraft-Diagrammen zufolge weist das transformierte Feld tatsächlich eine geringfügig niedrigere Vorhersagekraft als das originale Feld auf.

▶ Wählen Sie in der Ansicht "Felder" aus der Dropdown-Liste "Zu verwendende Version" in der Zeile reside die Option Original.

In Anbetracht der Tatsache, dass *Number of people in household* (Haushaltsgröße) in der Rangordnung der Felder nach Vorhersagekraft ziemlich weit unten steht, ist es unwahrscheinlich, dass dieser Punkt einen nennenswerten Effekt auf die Modellerstellung hat. Doch einer der Vorteile der interkativen Anwendung der automatisierten Datenaufbereitung ist es, dass die Transformationen ganz nach Bedarf fein abgestimmt werden können.

▶ Klicken Sie auf Ausführen, um die Transformationen anzuwenden.

Der Datensatz ist jetzt in der Hinsicht bereit für die Modellerstellung, dass die Rollen aller empfohlenen Einflussgrößen (sowohl neuer als auch alter) auf “Eingabe” gesetzt sind, wogegen die Rollen nicht empfohlener Einflussgrößen auf “Keine” gesetzt sind. Um einen Datensatz nur mit den empfohlenen Einflussgrößen zu erstellen, verwenden Sie die Einstellung “Transformationen anwenden” im Dialogfeld.

Automatische Verwendung der automatisierten Datenaufbereitung

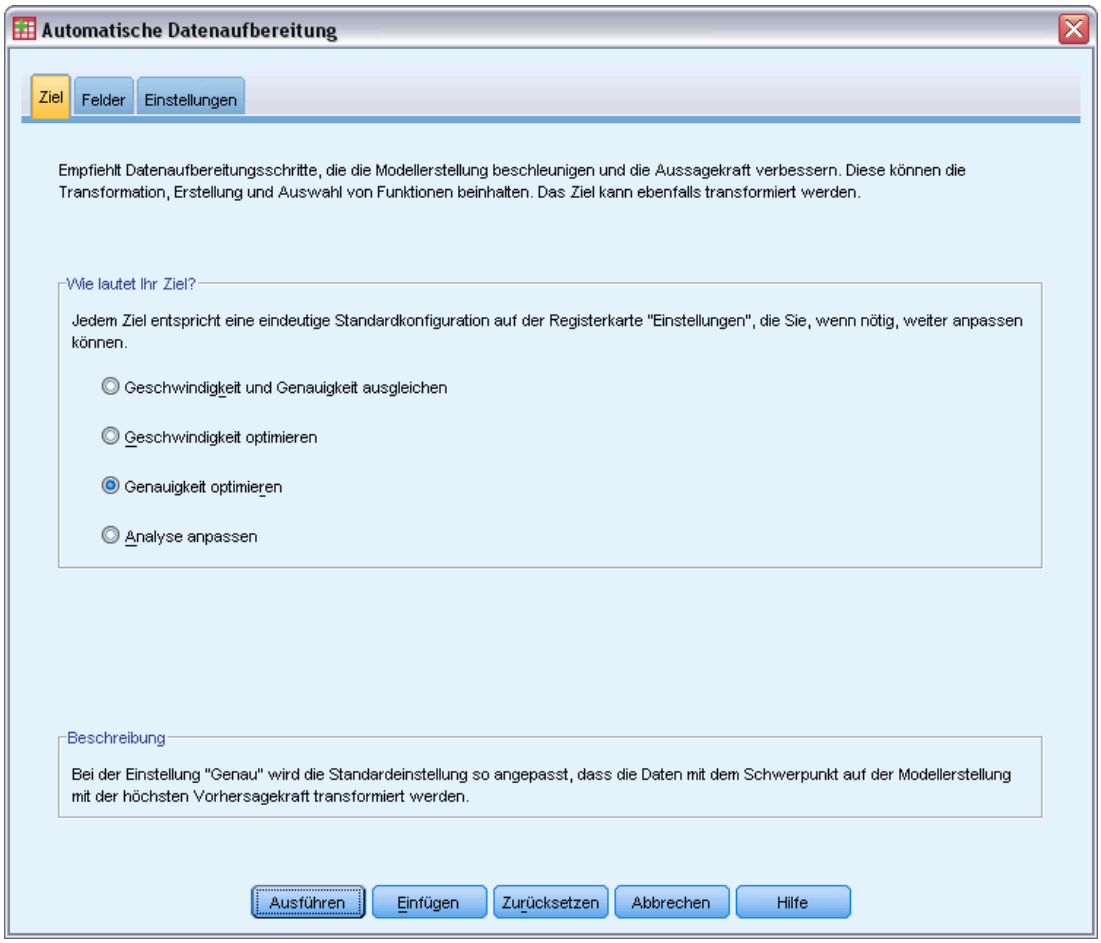
Eine Gruppe in der Kraftfahrzeugindustrie erfasst die Verkaufszahlen verschiedener Personenkraftwagen. Um starke und schwache Modelle identifizieren zu können, soll eine Beziehung zwischen den Fahrzeugverkaufszahlen und den Fahrzeugeigenschaften hergestellt werden. Diese Informationen sind in der Datei *car_sales_unprepared.sav* erfasst. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 139. Verwenden Sie die automatisierte Datenaufbereitung, um die Daten für die Analyse vorzubereiten. Erstellen Sie außerdem Modelle mit Daten “vor” und “nach” der Aufbereitung, um die Ergebnisse vergleichen zu können.

Vorbereitung der Daten

- Zur automatischen Ausführung der automatisierten Datenaufbereitung wählen Sie aus den Menüs:

  Transformieren
  Daten für Modellierung vorbereiten
  Automatisch...

Abbildung 8-13
Registerkarte "Ziel"

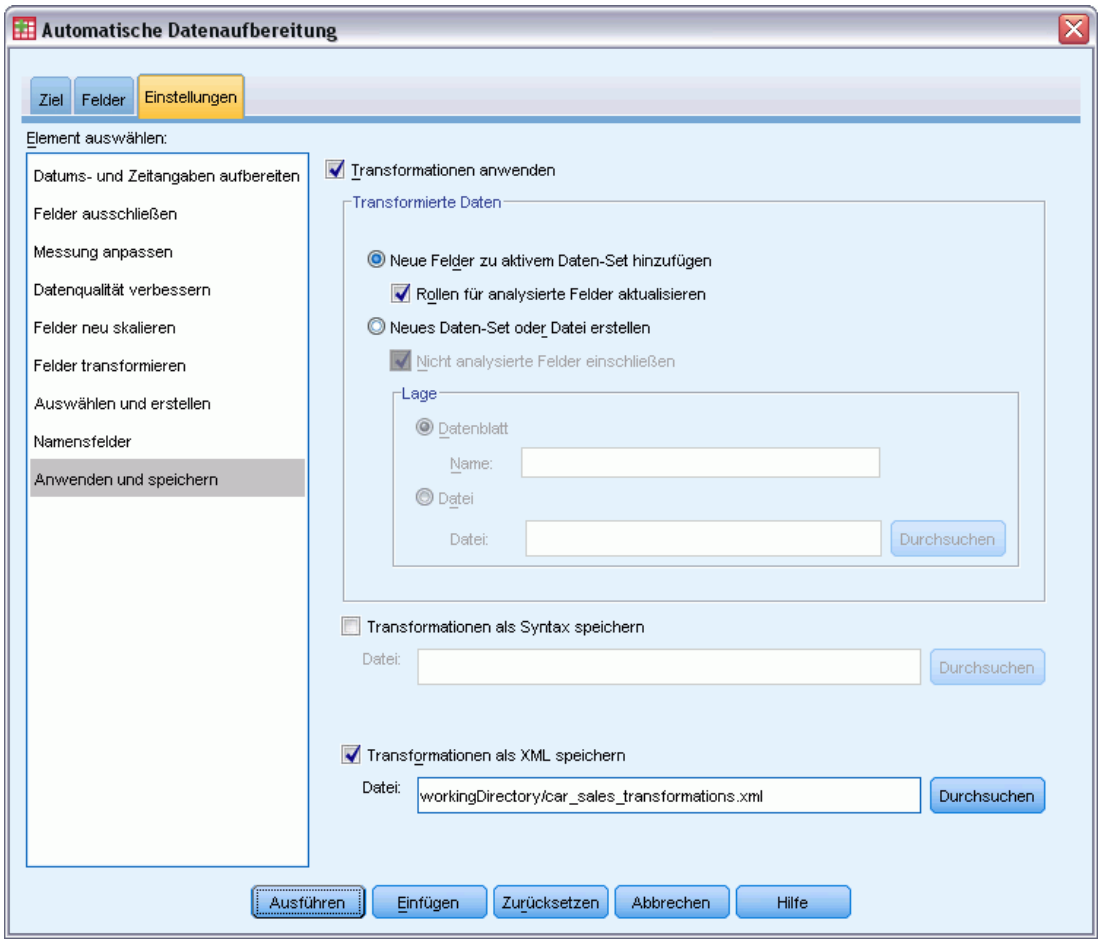

Automatische Datenaufbereitung
Ziel
Felder
Einstellungen
Empfiehlt Datenaufbereitungsschritte, die die Modellerstellung beschleunigen und die Aussagekraft verbessern. Diese können die Transformation, Erstellung und Auswahl von Funktionen beinhalten. Das Ziel kann ebenfalls transformiert werden.
Wie lautet Ihr Ziel?
Jedem Ziel entspricht eine eindeutige Standardkonfiguration auf der Registerkarte "Einstellungen", die Sie, wenn nötig, weiter anpassen können.
Geschwindigkeit und Genauigkeit ausgleichen
Geschwindigkeit optimieren
Genauigkeit optimieren
Analyse anpassen
Beschreibung
Bei der Einstellung "Genau" wird die Standardeinstellung so angepasst, dass die Daten mit dem Schwerpunkt auf der Modellerstellung mit der höchsten Vorhersagekraft transformiert werden.
Ausführen
Einfügen
Zurücksetzen
Abbrechen
Hilfe

- Wählen Sie Genauigkeit optimieren.

  Da das Zielfeld *Sales in thousands* (Verkäufe in Tausend) stetig ist und in der automatisierten Datenaufbereitung transformiert werden könnte, wollen Sie die Transformationen in einer XML-Datei speichern, damit Sie das Dialogfeld "Werte zurücktransformieren" verwenden können, um Vorhersagewerte des transformierten Ziels zurück auf ihre ursprüngliche Größe zu konvertieren.

- Klicken Sie auf die Registerkarte Einstellungen und anschließend auf die Einstellungen Anwenden und speichern.

Abbildung 8-14
Einstellungen "Anwenden und speichern"



- Wählen Sie Transformationen als XML speichern und klicken Sie auf Durchsuchen, um workingDirectory/car_sales_transformations.xml als den Pfad einzugeben, unter dem Sie die Datei für das Arbeitsverzeichnis speichern möchten.
- Klicken Sie auf Ausführen.

Diese Auswahl führt zu folgender Befehlssyntax:

```
*Automatic Data Preparation.
ADP
  /FIELDS TARGET=sales INPUT=resale type price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /PREPDATETIME DATEDURATION=YES(REFERENCE=YMD('2009-06-04') UNIT=AUTO)
    TIMEDURATION=YES(REFERENCE=HMS('08:43:35') UNIT=AUTO) EXTRACTYEAR=YES(SUFFIX='_year')
    EXTRACTMONTH=YES(SUFFIX='_month') EXTRACTDAY=YES(SUFFIX='_day')
    EXTRACTHOUR=YES(SUFFIX='_hour') EXTRACTMINUTE=YES(SUFFIX='_minute')
    EXTRACTSECOND=YES(SUFFIX='_second')
   /SCREENING PCTMISSING=YES(MAXPCT=50) UNIQUECAT=YES(MAXCAT=100) SINGLECAT=NO
  /ADJUSTLEVEL INPUT=YES TARGET=YES MAXVALORDINAL=10 MINVALCONTINUOUS=5
  /OUTLIERHANDLING INPUT=YES TARGET=NO CUTOFF=SD(3) REPLACEWITH=CUTOFFVALUE
  /REPLACEMISSING INPUT=YES TARGET=NO
  /REORDERNOMINAL INPUT=YES TARGET=NO
```

```
  /RESCALE INPUT=ZSCORE(MEAN=0 SD=1) TARGET=BOXCOX(MEAN=0 SD=1)
  /TRANSFORM MERGESUPERVISED=NO MERGEUNSUPERVISED=NO BINNING=NONE SELECTION=NO
    CONSTRUCTION=NO
  /CRITERIA SUFFIX(TARGET='_transformed' INPUT='_transformed')
  /OUTFILE PREPXML='/workingDirectory/car_sales_transformations.xml'.
TMS IMPORT
  /INFILE TRANSFORMATIONS='/workingDirectory/car_sales_transformations.xml'
    MODE=FORWARD (ROLES=UPDATE)
  /SAVE TRANSFORMED=YES.
EXECUTE.
```

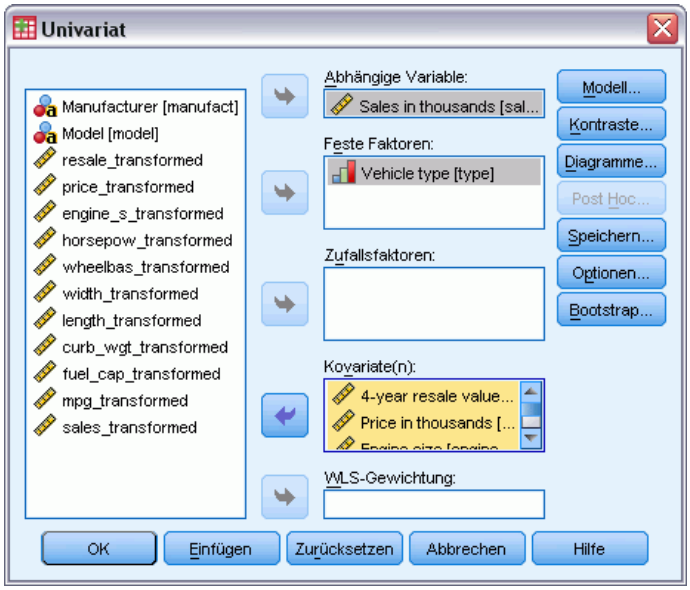
- Mithilfe des Befehls `ADP` werden das Zielfeld *sales* (Verkäufe) und die Eingabefelder *resale* (Wiederverkaufswert) durch *mpg* (Verbrauchswerte) aufbereitet.
- Der Unterbefehl `PREPDATETIME` wird aufgeführt, jedoch nicht angewendet, da keines der Felder ein Datums- oder ein Zeitfeld ist.
- Der Unterbefehl `ADJUSTLEVEL` wandelt Ordinalfelder mit über zehn Werten in stetige Felder und stetige Felder mit weniger als fünf Werten in Ordinalfelder.
- Der Unterbefehl `OUTLIERHANDLING` ersetzt Werte stetiger Eingaben (nicht das Ziel), die über drei Standardabweichungen vom Mittelwert entfernt sind, durch den Wert, der drei Standardabweichungen vom Mittelwert entfernt ist.
- Der Unterbefehl `REPLACEMISSING` ersetzt fehlende Eingabewerte (nicht das Ziel).
- Der Unterbefehl `REORDERNOMINAL` kodiert die Werte von nominalen Eingaben von “am seltensten auftretend” auf “am häufigsten auftretend” um.
- Der Unterbefehl `RESCALE` standardisiert stetige Eingaben mithilfe einer Z-Wert-Transformation auf einen Mittelwert von 0 und eine Standardabweichung von 1 und standardisiert das stetige Ziel mithilfe einer Box-Cox-Transformation auf einen Mittelwert von 0 und eine Standardabweichung von 1.
- Der Unterbfehl `TRANSFORM` deaktiviert alle von diesem Unterbefehl spezifizierten Standardvorgänge.
- Der Unterbefehl `CRITERIA` spezifiziert die Standardsuffixe für die Transformationen des Ziels und der Eingaben.
- Der Unterbefehl `OUTFILE` gibt an, dass die Transformationen unter */workingDirectory/car_sales_transformations.xml* gespeichert werden sollen, wobei */workingDirectory* der Pfad ist, unter dem Sie die Datei *car_sales_transformations.xml* speichern möchten.
- Der Befehl `TMS IMPORT` liest die Transformationen in *car_sales_transformations.xml* und wendet sie auf den aktiven Datensatz an, wobei die Rollen bestehender Felder, die transformiert werden, aktualisiert werden.
- Mit dem Befehl `EXECUTE` werden die Transformationen verarbeitet. Wenn Sie den Befehl `EXECUTE` als Teil eines längeren Syntaxstroms verwenden, können Sie ihn entfernen, um Verarbeitungszeit zu sparen.

Erstellen eines Modells mit unvorbereiteten Daten

- Zur Erstellung eines Modells mit den unvorbereiteten Daten wählen Sie aus den Menüs:

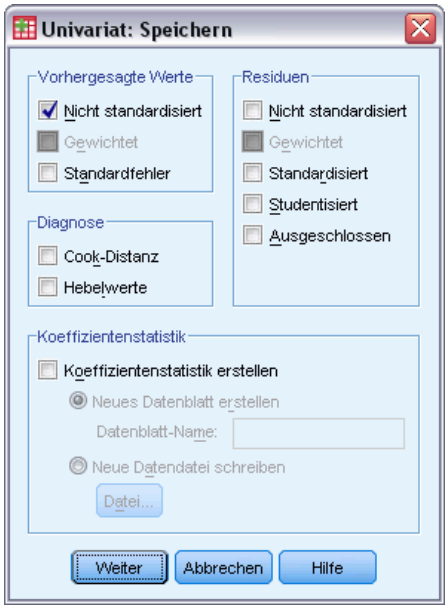
  Analysieren
  Allgemeines lineares Modell
  Univariat...

Abbildung 8-15
Dialogfeld "GLM-Univariat"



- Wählen Sie *Sales in thousands [sales]* (Verkäufe in Tausend [Verkaufszahlen]) als abhängige Variable aus.
- Wählen Sie *Vehicle type [type]* (Fahrzeugtyp [Typ]) als festen Faktor.
- Wählen Sie *4-year resale value [resale]* (Wiederverkaufswert 4 Jahre [Wiederverkauf]) durch *Fuel efficiency [mpg]* (Kraftstoffverwertung [Verbrauchswerte]) als Kovariaten aus.
- Klicken Sie auf Speichern.

Abbildung 8-16
Dialogfeld "Speichern"



- Wählen Sie in der Gruppe "Vorhersagewerte" die Option Nicht standardisiert.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "GLM - Univariat" auf OK.

Diese Auswahl führt zu folgender Befehlssyntax:

```
UNIANOVA sales BY type WITH resale price engine_s horsepow wheelbas width length
    curb_wgt fuel_cap mpg
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale price engine_s horsepow wheelbas width length curb_wgt fuel_cap
    mpg type.
```

Abbildung 8-17
Zwischensubjekteffekte für auf unvorbereiteten Daten basierte Modelle

Abhängige Variable:Sales in thousands

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	226123.658[a]	11	20556.696	5.050	.000
Konstanter Term	12227.688	1	12227.688	3.004	.086
resale	50.702	1	50.702	.012	.911
price	471.630	1	471.630	.116	.734
engine_s	19872.712	1	19872.712	4.882	.029
horsepow	9644.486	1	9644.486	2.369	.127
wheelbas	29824.272	1	29824.272	7.327	.008
width	263.465	1	263.465	.065	.800
length	1374.525	1	1374.525	.338	.562
curb_wgt	32762.692	1	32762.692	8.049	.005
fuel_cap	1124.237	1	1124.237	.276	.600
mpg	337.585	1	337.585	.083	.774
type	17668.779	1	17668.779	4.341	.040
Fehler	427402.183	105	4070.497		
Gesamt	1062354.955	117			
Korrigierte Gesamtvariation	653525.841	116			

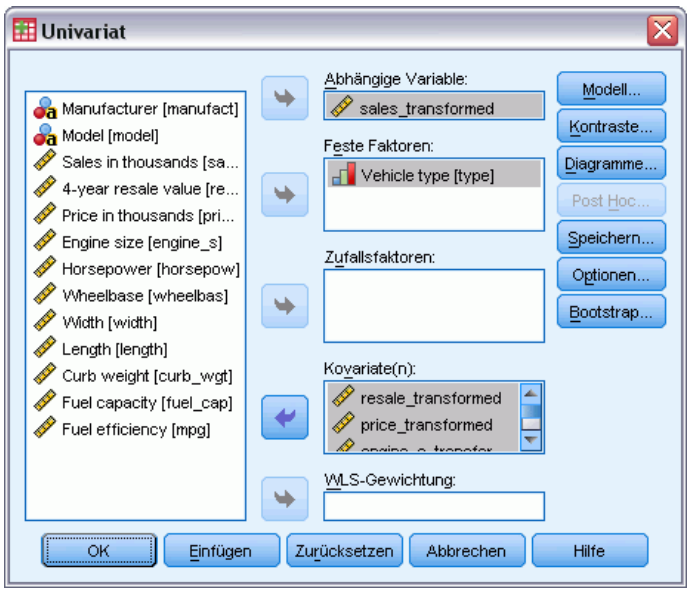
a. R-Quadrat = .346 (korrigiertes R-Quadrat = .277)

Die Standardausgabe für GLM-Univariat beinhaltet die Zwischensubjekteffekte, wobei es sich um eine Varianzanalyse-Tabelle handelt. Jeder Term in dem Modell sowie das Modell als Ganzes wird auf seine Fähigkeit getestet, Variationen in der abhängigen Variablen zu berücksichtigen. Hinweis: Variablenbezeichnungen sind in dieser Tabelle nicht dargestellt.

Die Einflussgrößen zeigen ein variierendes Signifikanzniveau; diejenigen, deren Signifikanzwerte kleiner als 0,05 sind, werden im Allgemeinen als für das Modell nützlich betrachtet.

Erstellen eines Modells mit den vorbereiteten Daten

Abbildung 8-18
Dialogfeld "GLM-Univariat"



- Zur Erstellung des Modells mit den vorbereiteten Daten rufen Sie das Dialogfeld "GLM-Univariat" auf.
- Deaktivieren Sie *Sales in thousands [sales]* (Verkäufe in Tausend []) und wählen Sie *sales_transformed* (Verkäufe_transformiert) als abhängige Variable aus.
- Deaktivieren Sie *4-year resale value [resale]* (Wiederverkaufswert 4 Jahre [Wiederverkauf]) durch *Fuel efficiency [mpg]* (Kraftstoffeffizienz [Verbrauchswerte]) und wählen Sie *resale_transformed* (Widerverkauf_transformiert) durch *mpg_transformed* (Verbrauchswerte_transformiert) als Kovariaten aus.
- Klicken Sie auf OK.

Diese Auswahl führt zu folgender Befehlssyntax:

```
UNIANOVA sales_transformed BY type WITH resale_transformed price_transformed
    engine_s_transformed horsepow_transformed wheelbas_transformed width_transformed
    length_transformed curb_wgt_transformed fuel_cap_transformed mpg_transformed
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /SAVE=PRED
  /CRITERIA=ALPHA(0.05)
  /DESIGN=resale_transformed price_transformed engine_s_transformed horsepow_transformed
    wheelbas_transformed width_transformed length_transformed curb_wgt_transformed
    fuel_cap_transformed mpg_transformed type.
```

Abbildung 8-19
Zwischensubjekteffekte für auf vorbereiteten Daten basierte Modelle

Abhängige Variable:sales_transformed

Quelle	Quadratsumme vom Typ III	df	Mittel der Quadrate	F	Sig.
Korrigiertes Modell	79.327[a]	11	7.212	13.638	.000
Konstanter Term	2.436	1	2.436	4.606	.034
resale_transformed	.954	1	.954	1.804	.181
price_transformed	9.271	1	9.271	17.533	.000
engine_s_transformed	2.885	1	2.885	5.456	.021
horsepow_transformed	.034	1	.034	.064	.801
wheelbas_transformed	1.213	1	1.213	2.293	.132
width_transformed	.037	1	.037	.071	.791
length_transformed	.265	1	.265	.501	.480
curb_wgt_transformed	.103	1	.103	.194	.660
fuel_cap_transformed	.132	1	.132	.249	.618
mpg_transformed	3.390	1	3.390	6.411	.012
type	4.007	1	4.007	7.579	.007
Fehler	76.673	145	.529		
Gesamt	156.000	157			
Korrigierte Gesamtvariation	156.000	156			

a. R-Quadrat = .509 (korrigiertes R-Quadrat = .471)

Zwischen dem auf den unvorbereiteten Daten erstellten Modell und dem auf den vorbereiteten Daten erstellten Modell gibt es einige interessante Unterschiede. So sei zunächst darauf hingewiesen, dass die Gesamtfreiheitsgrade zugenommen haben. Dies liegt an der Tatsache, dass fehlende Werte bei der automatisierten Datenaufbereitung durch abgeleitete Werte ersetzt wurden, so dass Datensätze, die bei dem ersten Modell listenweise entfernt wurden, beim zweiten Modell verfügbar sind. Noch beachtenswerter ist vielleicht, dass sich die Signifikanz bestimmter Einflussgrößen geändert hat. Zwar sind beide Modelle bei der Einschätzung identisch, dass die Motorgröße [*engine_s*] und der Fahrzeugtyp [*type*] für das Modell nützlich sind, doch sind der Radstand [*wheelbas*] und das Leergewicht [*curb_wgt*] nicht mehr signifikant, der Fahrzeugpreis [*price_transformed*] und die Kraftstoffverwertung [*mpg_transformed*] dagegen schon.

Woher kommt diese Veränderung? "Verkäufe" weist eine schiefe Verteilung auf, so dass der "Radstand" und das "Leergewicht" einige beeinflussende Datensätze umfasst haben könnten, die keinen Einfluss mehr hatten, als "Verkäufe" transformiert war. Eine andere Möglichkeit ist, dass die Zusatzfälle, die aufgrund fehlender Wertersetzung verfügbar sind, die statistische Signifikanz dieser Variablen verändert haben. In jedem Falle wären weitere Nachforschungen erforderlich, denen wir hier jedoch nicht nachgehen.

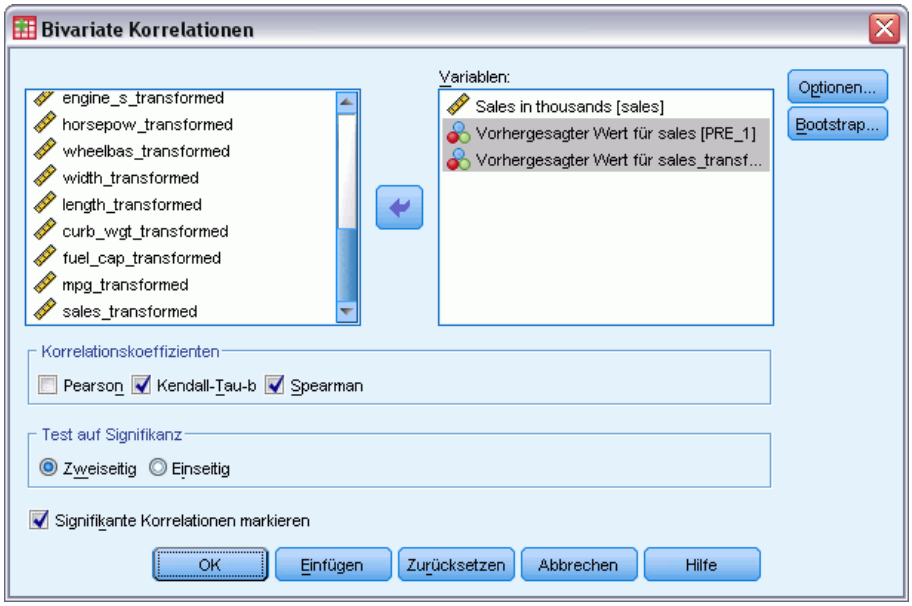
Hinweis: Sas Quadrat von R ist für das auf den vorbereiteten Daten erstellte Modell höher, doch da die Variable "Verkäufe" transformiert wurde, ist das nicht unbedingt der beste Maßstab für einen Vergleich der Qualität der Modelle. Stattdessen können Sie die nicht parametrischen Korrelationen zwischen den beobachteten Werten und den zwei Sätzen an Vorhersagewerten berechnen.

Vergleichen der Vorhersagewerte

▶ Für Korrelationen der Vorhersagewerte aus den zwei Modellen wählen Sie aus den Menüs:

Analysieren
  Korrelation
    Bivariat...

Abbildung 8-20
Dialogfeld "Bivariate Korrelationen"



▶ Wählen Sie *Sales in thousands [sales]* (Verkäufe in Tausend), *Predicted Value for sales [PRE_1]* (Vorhersagewert für Verkäufe) und *Predicted Values for sales_transformed [PRE_2]* (Vorhersagewerte für Verkäufe_transformiert) als Analysevariablen aus.

▶ Deaktivieren Sie Pearson und wählen Sie Kendall-Tau-b und Spearman in der Gruppe "Korrelationskoeffizienten".

Hinweis: *Predicted Values for sales_transformed [PRE_2]* (Vorhersagewerte für Verkäufe) kann für die Berechnung der nichtparametrischen Korrelationen verwendet werden, ohne auf die originale Größe zurücktransformiert werden zu müssen, da eine Rücktransformation die Rangordnung der Vorhersagewerte nicht ändert.

▶ Klicken Sie auf OK.

Diese Auswahl führt zu folgender Befehlssyntax:

```
NONPAR CORR
  /VARIABLES=sales PRE_1 PRE_2
  /PRINT=BOTH TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Abbildung 8-21
Nichtparametrische Korrelationen

			Sales in thousands	Vorhergesagter Wert für sales	Vorhergesagter Wert für sales_ transformed
Kendall-Tau-b	Sales in thousands	Korrelationskoeffizient	1.000	.376**	.484**
		Sig. (2-seitig)	.	.000	.000
		N	157	117	157
	Vorhergesagter Wert für sales	Korrelationskoeffizient	.376**	1.000	.655**
		Sig. (2-seitig)	.000	.	.000
		N	117	117	117
	Vorhergesagter Wert für sales_transformed	Korrelationskoeffizient	.484**	.655**	1.000
		Sig. (2-seitig)	.000	.000	.
		N	157	117	157
Spearman-Rho	Sales in thousands	Korrelationskoeffizient	1.000	.530**	.666**
		Sig. (2-seitig)	.	.000	.000
		N	157	117	157
	Vorhergesagter Wert für sales	Korrelationskoeffizient	.530**	1.000	.831**
		Sig. (2-seitig)	.000	.	.000
		N	117	117	117
	Vorhergesagter Wert für sales_transformed	Korrelationskoeffizient	.666**	.831**	1.000
		Sig. (2-seitig)	.000	.000	.
		N	157	117	157

**. Die Korrelation ist auf dem 0,01 Niveau signifikant (zweiseitig).

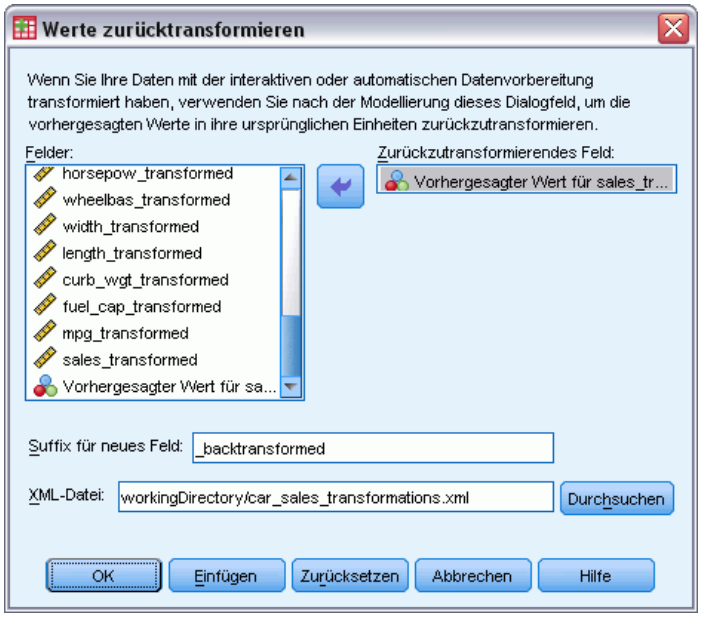
In der ersten Spalte ist zu sehen, dass die Vorhersagewerte für Modelle, die mit den vorbereiteten Daten erzeugt wurden, stärker mit den nach Kendall-Tau-b und Spearman-Rho beobachteten Werten korrelieren. Daraus lässt sich schließen, dass die Ausführung der automatisierten Datenaufbereitung das Modell verbessert hat.

Rücktransformieren der Vorhersagewerte

- Die vorbereiteten Daten umfassen eine Transformation von “Verkäufe”, so dass die Vorhersagewerte aus diesem Modell nicht direkt als Werte verwendet werden können. Zur Transformation der Vorhersagewerte auf die originale Größe wählen Sie aus den Menüs:

  Transformieren
  Daten für Modellierung vorbereiten
  Werte zurücktransformieren...

Abbildung 8-22
Dialogfeld "Werte zurücktransformieren"



- Wählen Sie *Predicted Value for sales_transformed [PRE_2]* (Vorhersagewerte für Verkäufe_transformiert) als zurückzutransformierendes Feld.
- Geben Sie _backtransformed als Suffix für das neue Feld ein.
- Geben Sie als Speicherort für die XML-Datei mit den Transformationen den Pfad workingDirectory\car_sales_transformations.xml ein, um die Datei im Arbeitsverzeichnis zu speichern.
- Klicken Sie auf OK.

Diese Auswahl führt zu folgender Befehlssyntax:

```
TMS IMPORT
  /INFILE TRANSFORMATIONS='workingDirectory/car_sales_transformations.xml'
    MODE=BACK (PREDICTED=PRE_2 SUFFIX='_backtransformed').
EXECUTE.
```

- Der Befehl `TMS IMPORT` liest die Transformationen in *car_sales_transformations.xml* und wendet die Rücktransformation auf *PRE_2* an.
- Das neue Feld, das die rücktransformierten Werte enthält, erhält die Bezeichnung *PRE_2_backtransformed*.
- Mit dem Befehl `EXECUTE` werden die Transformationen verarbeitet. Wenn Sie den Befehl `EXECUTE` als Teil eines längeren Syntaxstroms verwenden, können Sie ihn entfernen, um Verarbeitungszeit zu sparen.

Zusammenfassung

Mithilfe der automatisierten Datenaufbereitung erhalten Sie schnelle Datentransformationen, die Ihr Modell verbessern können. Wenn das Ziel transformiert ist, können Sie die Transformationen als XML-Datei speichern und das Dialogfeld “Werte zurücktransformieren” nutzen, um die Vorhersagewerte für das transformierte Ziel zurück auf die ursprüngliche Größe zu transformieren.

Kapitel
9

Ungewöhnliche Fälle identifizieren

Die Prozedur “Anomalie-Erkennung” sucht anhand von Abweichungen von den Normwerten der Gruppe nach ungewöhnlichen Fällen. Die Prozedur wurde für die Datenprüfung in der explorativen Datenanalyse konzipiert. Zweck der Prozedur ist das schnelle Erkennen von ungewöhnlichen Fällen, bevor mit anderen Analysen Schlüsse aus den Daten gezogen werden. Dieser Algorithmus dient der Erkennung von allgemeinen Anomalien. Dies bedeutet, dass sich die Definition eines anomalen Falls nicht auf eine bestimmte Anwendung beschränkt, bei der Anomalien sehr treffend definiert werden können, z. B. beim Erkennen von ungewöhnlichen Zahlungsmustern im Gesundheitswesen oder beim Aufdecken von Geldwäsche im Finanzwesen.

Algorithmus für “Ungewöhnliche Fälle identifizieren”

Dieser Algorithmus gliedert sich in drei Phasen:

Modellierung. Die Prozedur erstellt ein Clustermodell zum Ermitteln von natürlichen Gruppierungen (Clustern) innerhalb eines Daten-Sets, die andernfalls nicht erkennbar wären. Die Cluster beruhen auf einem Satz von Eingabevariablen. Das resultierende Clustermodell und ausreichende Statistiken zum Berechnen der Normwerte der Clustergruppen werden für die spätere Verwendung gespeichert.

Bewertung. Das Modell wird auf jeden Fall angewendet, um die Clustergruppe des Falls zu ermitteln. Dabei werden Indikatorvariablen für jeden Fall erstellt, um die Ungewöhnlichkeit jedes Falls in Bezug auf die entsprechende Clustergruppe zu messen. Die Fälle werden nach den Werten des Anomalie-Index sortiert. Der oberste Anteil der Fallliste stellt die Anomalien dar.

Argumentation. Für jeden anomalen Fall werden die Variablen nach den entsprechenden Variablenabweichungs-Indizes sortiert. Die obersten Variablen, deren Werte und die entsprechenden Normwerte werden als Gründe ausgegeben, warum ein Fall als Anomalie identifiziert wurde.

Identifizieren ungewöhnlicher Fälle in einer medizinischen Datenbank

Ein Analytiker, der mit der Erstellung von Prognosemodellen für die Ergebnisse von Schlaganfallbehandlungen betraut wurde, ist über die Qualität der Daten besorgt, weil solche Modelle bei ungewöhnlichen Beobachtungen anfällig sein können. Einige dieser Randbeobachtungen stellen wirklich einzigartige Fälle dar und eignen sich deswegen nicht für eine Vorhersage. Andere Beobachtungen stellen Dateneingabefehler dar, wobei die Werte technisch gesehen "richtig" sind und deswegen nicht mit Datenvalidierungsprozeduren abgefangen werden können.

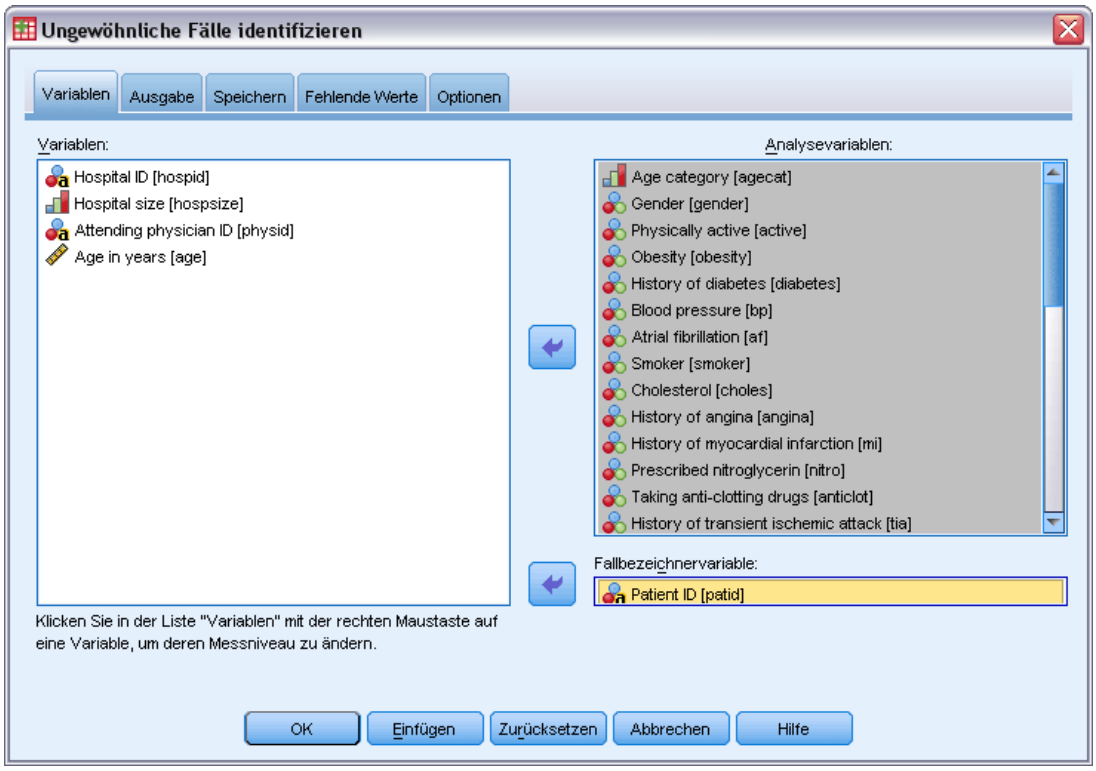
Diese Informationen finden Sie in der Datei *stroke_valid.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 139. Verwenden Sie die Prozedur "Ungewöhnliche Fälle identifizieren", um die Datendatei zu bereinigen. Syntax, mit denen Sie diese Analysen nachvollziehen können, befindet sich in der Datei *detectanomaly_stroke.sps*.

Durchführen der Analyse

- Um ungewöhnliche Fälle zu identifizieren, wählen Sie die folgenden Befehle aus den Menüs aus:

  Daten
  Ungewöhnliche Fälle identifizieren...

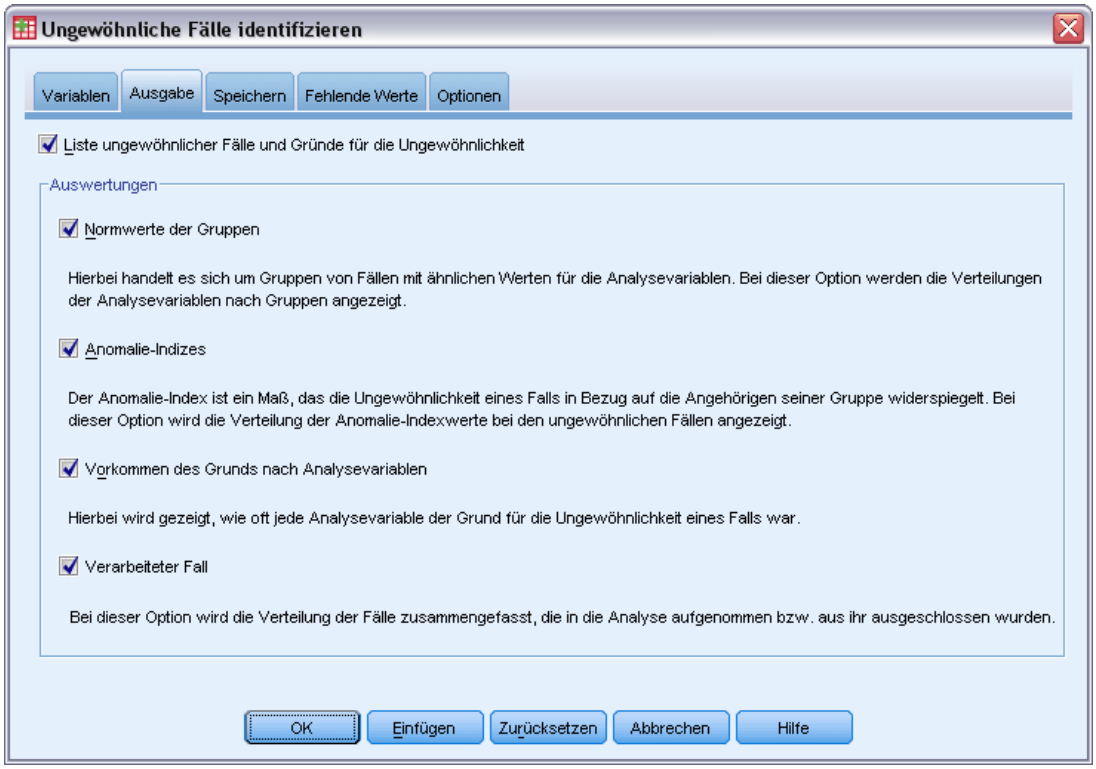
Abbildung 9-1
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Variablen"



- Wählen Sie die Variablen von *Age category* bis *Stroke between 3 and 6 months* als Analysevariablen aus.

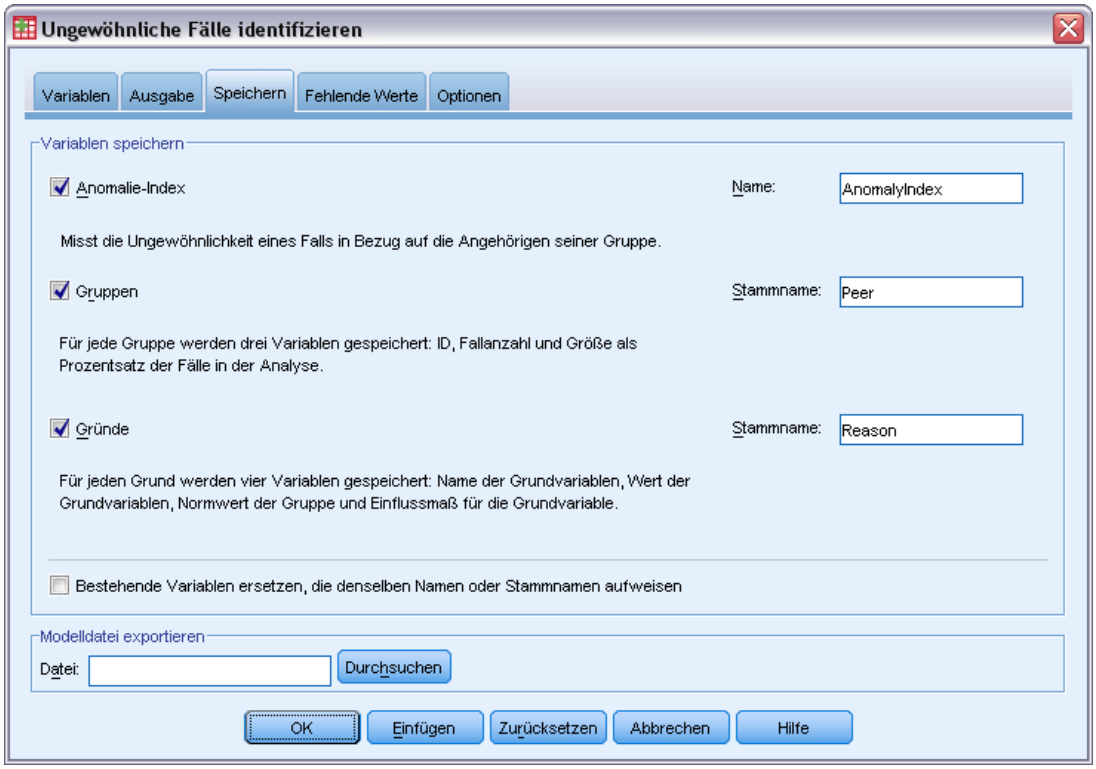
- Wählen Sie *Patient ID* als Fallbezeichnervariable aus.
- Klicken Sie auf die Registerkarte Ausgabe.

Abbildung 9-2
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Ausgabe"


Ungewöhnliche Fälle identifizieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Liste ungewöhnlicher Fälle und Gründe für die Ungewöhnlichkeit
Auswertungen
Normwerte der Gruppen
Hierbei handelt es sich um Gruppen von Fällen mit ähnlichen Werten für die Analysevariablen. Bei dieser Option werden die Verteilungen der Analysevariablen nach Gruppen angezeigt.
Anomalie-Indizes
Der Anomalie-Index ist ein Maß, das die Ungewöhnlichkeit eines Falls in Bezug auf die Angehörigen seiner Gruppe widerspiegelt. Bei dieser Option wird die Verteilung der Anomalie-Indexwerte bei den ungewöhnlichen Fällen angezeigt.
Vorkommen des Grunds nach Analysevariablen
Hierbei wird gezeigt, wie oft jede Analysevariable der Grund für die Ungewöhnlichkeit eines Falls war.
Verarbeiteter Fall
Bei dieser Option wird die Verteilung der Fälle zusammengefasst, die in die Analyse aufgenommen bzw. aus ihr ausgeschlossen wurden.
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

- Wählen Sie Normwerte der Gruppen, Anomalie-Indizes, Vorkommen des Grunds nach Analysevariablen und Verarbeitete Fälle aus.
- Klicken Sie auf die Registerkarte Speichern.

Abbildung 9-3
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Speichern"

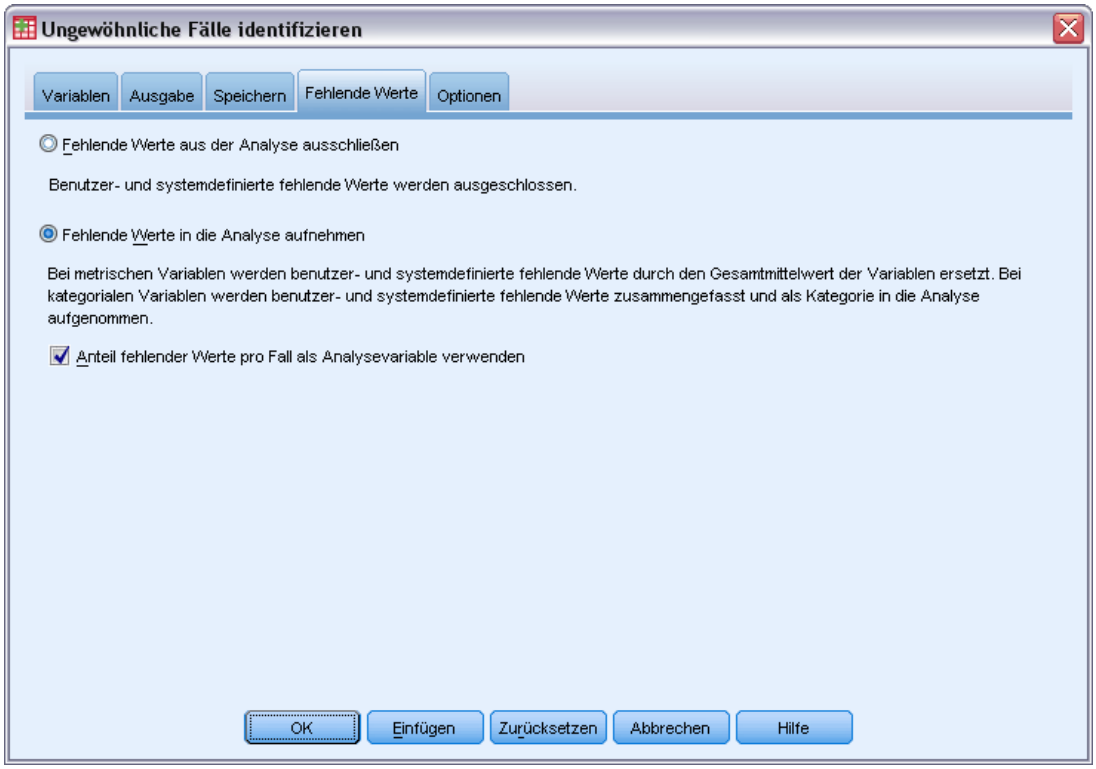


▶ Wählen Sie Anomalie-Index, Gruppen und Gründe aus.

Wenn Sie diese Ergebnisse speichern, können Sie ein sinnvolles Streudiagramm erstellen, mit dem die Ergebnisse zusammengefasst werden.

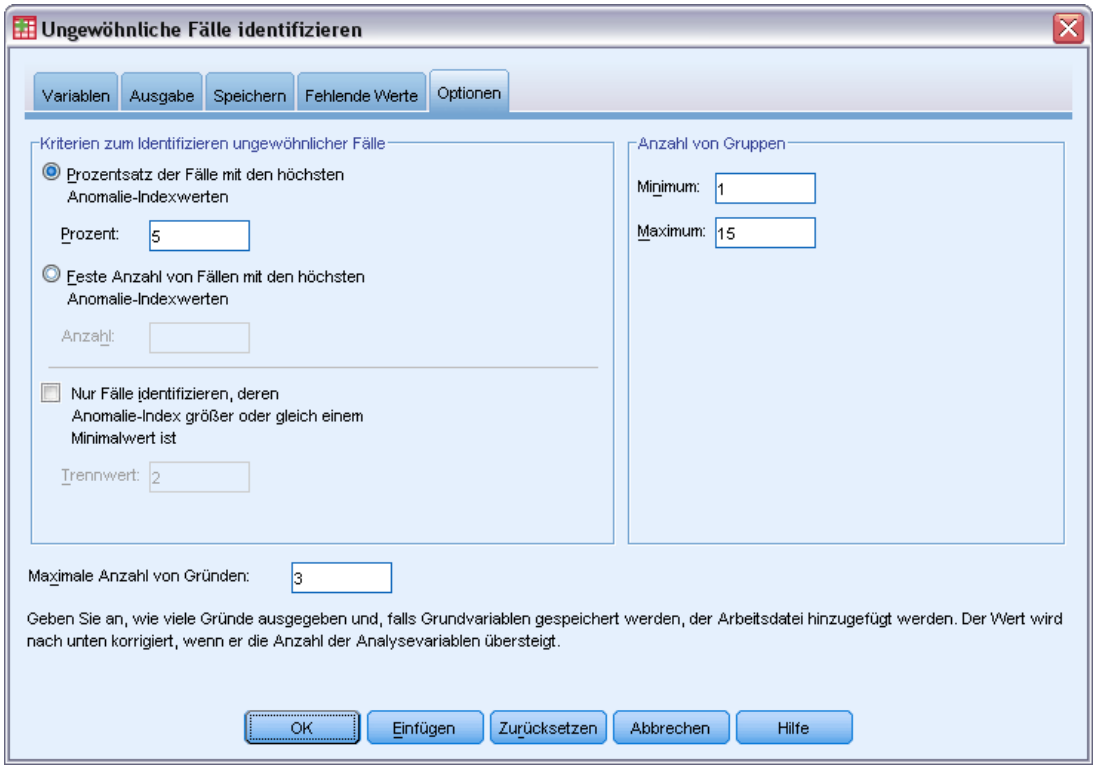
▶ Klicken Sie auf die Registerkarte Fehlende Werte.

Abbildung 9-4
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Fehlende Werte"


Ungewöhnliche Fälle identifizieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Fehlende Werte aus der Analyse ausschließen
Benutzer- und systemdefinierte fehlende Werte werden ausgeschlossen.
Fehlende Werte in die Analyse aufnehmen
Bei metrischen Variablen werden benutzer- und systemdefinierte fehlende Werte durch den Gesamtmittelwert der Variablen ersetzt. Bei kategorialen Variablen werden benutzer- und systemdefinierte fehlende Werte zusammengefasst und als Kategorie in die Analyse aufgenommen.
Anteil fehlender Werte pro Fall als Analysevariable verwenden
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

- Wählen Sie Fehlende Werte in die Analyse aufnehmen aus. Dies ist notwendig, weil viele benutzerdefinierte fehlende Werte für Patienten vorliegen, die vor oder während der Behandlung gestorben sind. Der Analyse wird eine zusätzliche metrische Variable hinzugefügt, mit der der Anteil der fehlenden Werte pro Fall aufgezeichnet wird.
- Klicken Sie auf die Registerkarte Optionen.

Abbildung 9-5
Dialogfeld "Ungewöhnliche Fälle identifizieren", Registerkarte "Optionen"


Ungewöhnliche Fälle identifizieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Kriterien zum Identifizieren ungewöhnlicher Fälle
Prozentsatz der Fälle mit den höchsten Anomalie-Indexwerten
Prozent: 5
Feste Anzahl von Fällen mit den höchsten Anomalie-Indexwerten
Anzahl:
Nur Fälle identifizieren, deren Anomalie-Index größer oder gleich einem Minimalwert ist
Trennwert: 2
Anzahl von Gruppen
Minimum: 1
Maximum: 15
Maximale Anzahl von Gründen: 3
Geben Sie an, wie viele Gründe ausgegeben und, falls Grundvariablen gespeichert werden, der Arbeitsdatei hinzugefügt werden. Der Wert wird nach unten korrigiert, wenn er die Anzahl der Analysevariablen übersteigt.
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

- Geben Sie als Prozentsatz der Fälle, die als anomal betrachtet werden sollen, den Wert 2 ein.
- Deaktivieren Sie Nur Fälle identifizieren, deren Anomalie-Index größer oder gleich einem Minimalwert ist.
- Geben Sie als maximale Anzahl von Gründen den Wert 3 ein.
- Klicken Sie auf OK.

Zusammenfassung der Fallverarbeitung

Abbildung 9-6
Zusammenfassung der Fallverarbeitung

		N	% vonkombiniert	% von gesamt
Gruppen-ID	1	710	67.7%	67.7%
	2	90	8.6%	8.6%
	3	248	23.7%	23.7%
Kombiniert		1048	100.0%	100.0%
Gesamt		1048		100.0%

Jeder Fall wird in eine Gruppe ähnlicher Fälle aufgenommen. Die Zusammenfassung der Fallverarbeitung zeigt, wie viele Gruppen erstellt wurden, sowie die Anzahl und den Prozentsatz von Fällen in jeder Gruppe.

Liste der Indizes anomaler Fälle

Abbildung 9-7
Liste der Indizes anomaler Fälle

Fall	patid	Anomaly Index
843	7840326167	2.837
510	0714726620	2.022
623	6553808330	2.014
501	6461046805	2.002
607	1077125669	1.897
884	2260043998	1.889
614	4030164769	1.869
241	1038840465	1.865
13	2191527525	1.826
172	4458028382	1.786
705	1336411777	1.778
651	4103977868	1.767
384	2247641363	1.767
839	0437454972	1.766
861	9746101913	1.757
19	7237535360	1.756
806	4391632997	1.756
871	6961938294	1.739
239	7315965190	1.738
887	6044244232	1.737
245	0816869249	1.736

Der Anomalie-Index ist ein Maß, das die Ungewöhnlichkeit eines Falls in Bezug auf die Angehörigen seiner Gruppe widerspiegelt. Dabei werden die 2 % der Fälle mit den höchsten Werten des Anomalie-Index sowie deren Fallnummern und Bezeichner angezeigt. Es werden 21 Fälle mit Werten von 1,736 bis 2,837 ausgegeben. Es liegt ein relativ großer Unterschied zwischen dem Wert des Anomalie-Index des ersten und des zweiten Falls in der Liste vor. Dies legt nahe, dass Fall 843 wahrscheinlich anomal ist. Die anderen Fälle müssen einzeln beurteilt werden.

Liste der Gruppen-IDs anomaler Fälle

Abbildung 9-8
Liste der Gruppen-IDs anomaler Fälle

Fall	patid	Gruppen-ID	Gruppengröße	Gruppengröße in Prozent
843	7840326167	3	248	23.7%
510	0714726620	3	248	23.7%
623	6553808330	3	248	23.7%
501	6461046805	3	248	23.7%
607	1077125669	3	248	23.7%
884	2260043998	3	248	23.7%
614	4030164769	3	248	23.7%
241	1038840465	3	248	23.7%
13	2191527525	3	248	23.7%
172	4458028382	3	248	23.7%
705	1336411777	1	710	67.7%
651	4103977868	1	710	67.7%
384	2247641363	3	248	23.7%
839	0437454972	3	248	23.7%
861	9746101913	3	248	23.7%
19	7237535360	1	710	67.7%
806	4391632997	1	710	67.7%
871	6961938294	1	710	67.7%
239	7315965190	3	248	23.7%
887	6044244232	1	710	67.7%
245	0816869249	3	248	23.7%

Die potenziell anomalen Fälle werden zusammen mit Informationen zu deren Gruppenmitgliedschaft angezeigt. Die ersten 10 Fälle (und insgesamt 15 Fälle) gehören zu Gruppe 3; alle weiteren zu Gruppe 1.

Liste der Gründe anomaler Fälle

Abbildung 9-9
Liste der Gründe anomaler Fälle

Grund: 1

Fall	patid	Grundvariable	Variablenbeeinflussung	Variablenwert	Normwert der Variablen
581	7516953949	physid	,081	176466	828754
841	7469179281	physid	,094	237547	828754
497	8879591858	physid	,073	037350	828754
524	6395130127	rankin3	,094	1	0
432	9064917517	rankin1	,094	2	0
815	9741176885	physid	,087	995409	828754
1014	9353251878	physid	,091	185703	828754
658	8918339607	barthel1	,123	60	95
313	1368252467	barthel2	,088	80	100
934	0621567299	physid	,104	680253	828754
362	9355732120	rankin2	,091	1	0
385	2554580988	rankin2	,092	1	0
966	4971530904	physid	,106	249058	828754
198	6240985380	rankin1	,101	2	0
181	7311392948	barthel1	,126	60	95
365	3548308139	physid	,087	993921	828754

Die Grundvariablen sind die Variablen, die am meisten dazu beitragen, dass ein Fall als ungewöhnlich eingestuft wird. Für jeden anomalen Fall werden die primäre Grundvariable, deren Einflussmaß und deren Wert für den Fall sowie der Normwert der Gruppe angezeigt. Wenn bei einer kategorialen Variablen als Normwert für die Gruppe *(Fehlender Wert)* angegeben ist, weist die Mehrzahl der Fälle in der Gruppe einen fehlenden Wert für diese Variable auf.

Das Einflussmaß der Variable ist der proportionale Beitrag der Grundvariable zur Abweichung des Falls von seiner Gruppe. Es liegen 38 Analysevariablen vor (einschließlich der Variablen für den fehlenden Anteil). Das erwartete Einflussmaß einer Variablen beträgt daher 1/38 = 0,026. Das Einflussmaß der Variable *cost* für Fall 843 beträgt 0,411, was relativ gesehen groß ist. Der Wert von *cost* für Fall 843 ist 200,51; der Durchschnitt für die Fälle in Gruppe 3 ist 19,83.

Im Dialogfeld wurde festgelegt, dass Ergebnisse für die ersten drei Gründe ausgegeben werden sollen.

- ► Um die Ergebnisse für die anderen Gründe anzuzeigen, doppelklicken Sie auf die Tabelle.
- ► Verschieben Sie *Grund* aus der Schichtendimension in die Zeilendimension.

Abbildung 9-10
Liste der Gründe anomaler Fälle (die ersten 8 Fälle)

Fall	Grund	patid	Grund Variablen	Variablen einflussung	Variablen wert	Normwert der Variablen
843	1	7840326167	cost	.411	200.51	19.83
	2	7840326167	barthel1	.076	65	(Missing Value)
	3	7840326167	rankin1	.044	2	(Missing Value)
510	1	0714726620	cost	.120	96.59	19.83
	2	0714726620	barthel1	.083	80	(Missing Value)
	3	0714726620	rehab	.068	3	(Missing Value)
623	1	6553808330	cost	.175	114.01	19.83
	2	6553808330	surgery	.089	2	(Missing Value)
	3	6553808330	barthel1	.089	70	(Missing Value)
501	1	6461046805	barthel1	.084	80	(Missing Value)
	2	6461046805	rehab	.068	3	(Missing Value)
	3	6461046805	rankin1	.063	1	(Missing Value)
607	1	1077125669	cost	.126	96.11	19.83
	2	1077125669	barthel1	.094	85	(Missing Value)
	3	1077125669	rehab	.072	3	(Missing Value)
884	1	2260043998	cost	.138	99.73	19.83
	2	2260043998	barthel1	.114	65	(Missing Value)
	3	2260043998	rehab	.072	3	(Missing Value)
614	1	4030164769	barthel1	.085	45	(Missing Value)
	2	4030164769	rankin1	.085	3	(Missing Value)
	3	4030164769	recbart1	.062	2	(Missing Value)

Bei dieser Einstellung ist es einfach, die relativen Beiträge der ersten drei Gründe für jeden Fall zu vergleichen. Wie vermutet, wird Fall 843 als anomal betrachtet, weil *cost* für diesen Fall einen ungewöhnlich hohen Wert aufweist. Im Gegensatz dazu trägt kein einzelner Grund mehr als 0,10 zur Ungewöhnlichkeit von Fall 501 bei.

Normwerte der metrischen Variablen

Abbildung 9-11
Normwerte der metrischen Variablen

		Gruppen-ID			Kombiniert
		1	2	3	
Length of stay for rehabilitation	Mean	16.55	16.39	15.91	16.39
	Std. Deviation	12.596	.000	6.834	10.887
Total treatment and rehabilitation costs in thousands	Mean	42.4673	3.5089	19.8273	33.7641
	Std. Deviation	26.45401	.50997	20.17309	27.31266
Missing Proportion	Mean	.006	.541	.354	.134
	Std. Deviation	.021	2.9E-016	.083	.197

Die Liste mit den Normwerten der metrischen Variablen enthält den Mittelwert und die Standardabweichung jeder Variablen pro Gruppe und insgesamt. Bei einem Vergleich der Gruppen finden Sie Hinweise darauf, welche Variablen zum Bilden der Gruppen beitragen.

So weist der Mittelwert von *Length of stay for rehabilitation* beispielsweise in allen drei Gruppen ähnliche Werte auf. Dies bedeutet, dass die Variable nicht zum Bilden der Gruppen beiträgt. Sie können jedoch die Variablen *Total treatment and rehabilitation costs in thousands* und *Missing Proportion* nutzen, um Näheres über die Gruppenmitgliedschaften zu erfahren.

Gruppe 1 weist die höchste mittleren Kosten und die wenigsten fehlenden Werte auf. Gruppe 2 zeichnet sich durch sehr niedrige Kosten und viele fehlende Werte aus. In Gruppe 3 finden sich mittlere Kosten und mäßig viele fehlende Werte.

Dies deutet darauf hin, dass Gruppe 2 aus Patienten besteht, die bereits bei der Ankunft verstorben waren. Daher fielen niedrige Kosten an, und alle Behandlungs- und Rehabilitationsvariablen weisen fehlende Werte auf. Gruppe 3 enthält wahrscheinlich viele Patienten, die bei der Behandlung starben. Daher fielen Behandlungskosten an, aber keine Rehabilitationskosten, und die Rehabilitationsvariablen weisen fehlende Werte auf. Gruppe 1 besteht wahrscheinlich fast ausschließlich aus Patienten, die die Behandlung und die Rehabilitation überlebt haben. Dadurch fielen die höchsten Kosten an.

Normwerte der kategorialen Variablen

Abbildung 9-12
Normwerte der kategorialen Variablen (die ersten 10 Variablen)

		Gruppen-ID			
		1	2	3	Kombiniert
Age category	Häufigste Kategorie	2	3	2	2
	Häufigkeit	277	25	81	383
	Prozent	39.0%	27.8%	32.7%	36.5%
Gender	Häufigste Kategorie	0	0	1	0
	Häufigkeit	361	46	126	529
	Prozent	50.8%	51.1%	50.8%	50.5%
Physically active	Häufigste Kategorie	1	0	0	0
	Häufigkeit	373	55	139	531
	Prozent	52.5%	61.1%	56.0%	50.7%
Obesity	Häufigste Kategorie	0	0	0	0
	Häufigkeit	555	67	178	800
	Prozent	78.2%	74.4%	71.8%	76.3%
History of diabetes	Häufigste Kategorie	0	0	0	0
	Häufigkeit	665	80	219	964
	Prozent	93.7%	88.9%	88.3%	92.0%
Blood pressure	Häufigste Kategorie	1	1	1	1
	Häufigkeit	445	49	139	633
	Prozent	62.7%	54.4%	56.0%	60.4%
Atrial fibrillation	Häufigste Kategorie	0	0	0	0
	Häufigkeit	641	83	216	940
	Prozent	90.3%	92.2%	87.1%	89.7%
Smoker	Häufigste Kategorie	0	0	0	0
	Häufigkeit	578	69	179	826
	Prozent	81.4%	76.7%	72.2%	78.8%
Cholesterol	Häufigste Kategorie	0	0	0	0
	Häufigkeit	406	52	136	594
	Prozent	57.2%	57.8%	54.8%	56.7%
History of angina	Häufigste Kategorie	0	0	0	0
	Häufigkeit	493	52	167	712
	Prozent	69.4%	57.8%	67.3%	67.9%

Die Normwerte der kategorialen Variablen dienen demselben Zweck wie die Normwerte der metrischen Variablen. Bei den Normwerten der kategorialen Variablen werden jedoch die häufigste Kategorie sowie die Anzahl und der Prozentsatz an Fällen in der Gruppe ausgegeben, die in diese Kategorie fallen. Ein Vergleich der Werte ist etwas komplizierter. So kann es beispielsweise auf den ersten Blick scheinen, dass *Gender* mehr zum Bilden der Gruppen beiträgt als *Smoker*, weil die häufigste Kategorie für *Smoker* in allen drei Gruppen dieselbe ist,

die häufigste Kategorie für *Gender* in Gruppe 3 jedoch abweicht. Da *Gender* aber nur zwei Werte annehmen kann, können Sie schlussfolgern, dass 49,2 % der Fälle in Gruppe 3 den Wert 0 aufweisen. Dies ähnelt stark den Prozentsätzen in den anderen Gruppen. Im Gegensatz dazu variieren die Prozentsätze für *Smoker* zwischen 72,2 % und 81,4 %.

Abbildung 9-13
Normwerte der kategorialen Variablen (ausgewählte Variablen)

		Gruppen-ID			
		1	2	3	Kombiniert
Dead on arrival	Häufigste Kategorie	0	1	0	0
	Häufigkeit	710	90	248	958
	Prozent	100.0%	100.0%	100.0%	91.4%
Initial Rankin score	Häufigste Kategorie	0	(Missing Value)	5	5
	Häufigkeit	166	90	104	193
	Prozent	23.4%	100.0%	41.9%	18.4%
CAT scan result	Häufigste Kategorie	0	(Missing Value)	0	0
	Häufigkeit	607	90	184	791
	Prozent	85.5%	100.0%	74.2%	75.5%
Clot-dissolving drugs	Häufigste Kategorie	2	(Missing Value)	0	2
	Häufigkeit	318	90	129	394
	Prozent	44.8%	100.0%	52.0%	37.6%
Died in hospital	Häufigste Kategorie	0	(Missing Value)	1	0
	Häufigkeit	710	90	171	787
	Prozent	100.0%	100.0%	69.0%	75.1%
Treatment result	Häufigste Kategorie	1	(Missing Value)	1	1
	Häufigkeit	524	90	96	620
	Prozent	73.8%	100.0%	38.7%	59.2%
Post-event preventative surgery	Häufigste Kategorie	0	(Missing Value)	(Missing Value)	0
	Häufigkeit	323	90	171	369
	Prozent	45.5%	100.0%	69.0%	35.2%
Post-event rehabilitation	Häufigste Kategorie	0	(Missing Value)	(Missing Value)	0
	Häufigkeit	278	90	171	314
	Prozent	39.2%	100.0%	69.0%	30.0%

Die durch die Normwerte der metrischen Variablen nahe gelegte Vermutung bestätigt sich im unteren Teil der Tabelle mit den Normwerten der kategorialen Variablen. Gruppe 2 besteht vollständig aus Patienten, die bereits bei der Ankunft verstorben waren. Deshalb fehlen alle Werte der Behandlungs- und Rehabilitationsvariablen. Die meisten Patienten in Gruppe 3 (69,0%) starben während der Behandlung. Daher ist die häufigste Kategorie für die Rehabilitationsvariablen *(Fehlender Wert)*.

Auswertung des Anomalie-Index

Abbildung 9-14
Auswertung des Anomalie-Index

	Anzahl anomaler Fälle	Minimum	Maximum	Mittelwert	Std. Deviation
Anomalie-Index	24	1,322	1,550	1,387	,068

Die Anzahl anomaler Fälle wird folgendermaßen bestimmt: Der Prozentsatz anomaler Fälle ist 2%.

Diese Tabelle enthält Auswertungsstatistiken für die Werte des Anomalie-Index von Fällen in der Anomalie-Liste.

Auswertung der Gründe

Abbildung 9-15
Auswertung der Gründe (Behandlungs- und Rehabilitationsvariablen)

	Auftreten als Grund		Statistiken der Variablenbeeinflussung			
	Häufigkeit	Prozent	Minimum	Maximum	Mittelwert	Std. Deviation
Dead on arrival	0	,0%	.	.	.	.
Initial Rankin score	0	,0%	.	.	.	.
CAT scan result	0	,0%	.	.	.	.
Clot-dissolving drugs	0	,0%	.	.	.	.
Died in hospital	0	,0%	.	.	.	.
Treatment result	1	4,2%	,110	,110	,110	.
Post-event preventative surgery	0	,0%	.	.	.	.
Post-event rehabilitation	0	,0%	.	.	.	.
Rankin score at 1 month	0	,0%	.	.	.	.
Rankin score at 3 months	0	,0%	.	.	.	.
Rankin score at 6 months	7	29,2%	,074	,080	,076	,003
Barthel index at 1 month	6	25,0%	,075	,136	,109	,021
Barthel index at 3 months	4	16,7%	,084	,118	,100	,018
Barthel index at 6 months	5	20,8%	,093	,108	,098	,006
Recoded Barthel index at 1 month	0	,0%	.	.	.	.
Recoded Barthel index at 3 months	0	,0%	.	.	.	.
Recoded Barthel index at 6 months	0	,0%	.	.	.	.
Length of stay for rehabilitation	0	,0%	.	.	.	.
Total treatment and rehabilitation costs in thousands	0	,0%	.	.	.	.
Anteil fehlend	0	,0%	.	.	.	.
Insgesamt	24	100,0%	,069	,136	,094	,019

In dieser Tabelle wird jede Analysevariable im Hinblick auf ihre Rolle als primärer Grund ausgewertet. Die meisten Variablen, z. B. *Dead on arrival* bis *Post-event rehabilitation* sind keine primären Gründe für die Fälle in der Anomalie-Liste. *Barthel index at 1 month* ist der häufigste Grund, *Total treatment and rehabilitation costs in thousands* der zweithäufigste. Es werden die Einflussstatistiken der Variablen ausgewertet. Dabei werden für jede Variable der kleinste,

größte und mittlere Einfluss sowie bei Variablen, die bei mehr als einem Fall die Grundvariablen sind, die Standardabweichung ausgegeben.

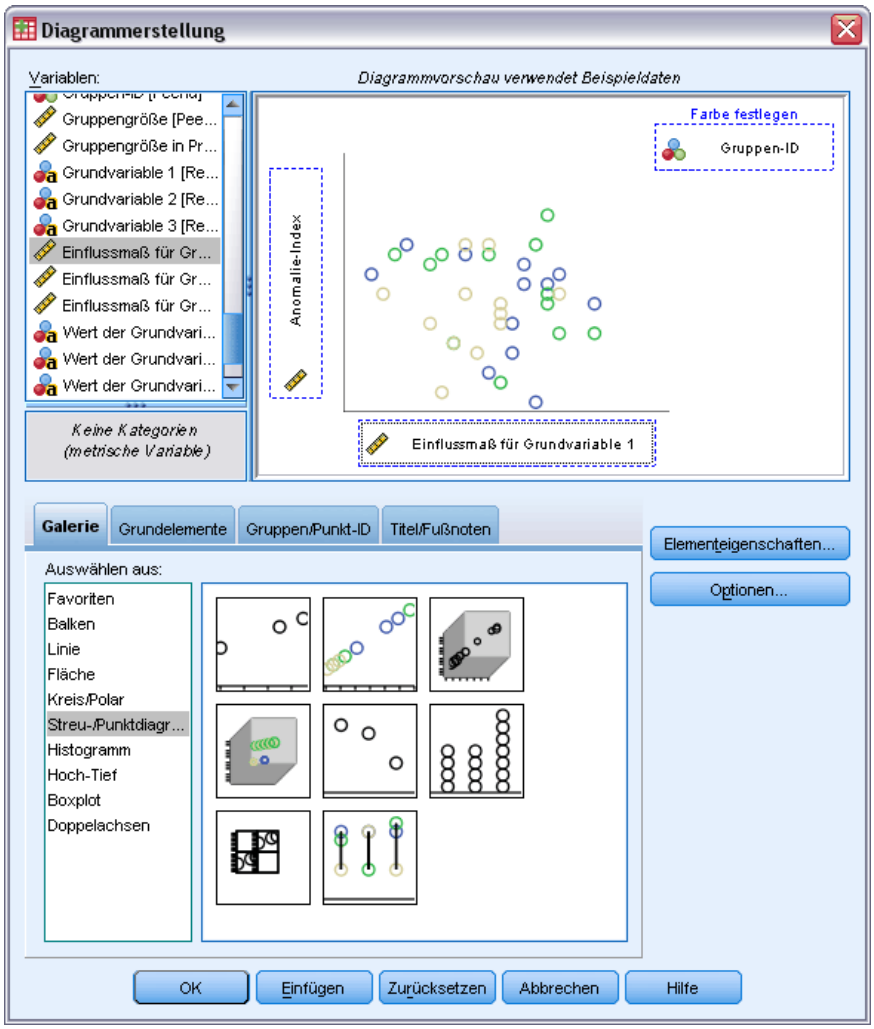
Streudiagramm des Anomalie-Index über den Variableneinfluss

Die Tabellen enthalten viele nützliche Informationen. Es kann jedoch schwierig sein, die wechselseitigen Beziehungen zu erfassen. Mit den gespeicherten Variablen können Sie eine Grafik erstellen, die Ihnen diese Aufgabe erleichtert.

- Um dieses Streudiagramm zu erstellen, wählen Sie die folgenden Befehle aus den Menüs aus:

Grafiken
  Diagrammerstellung...

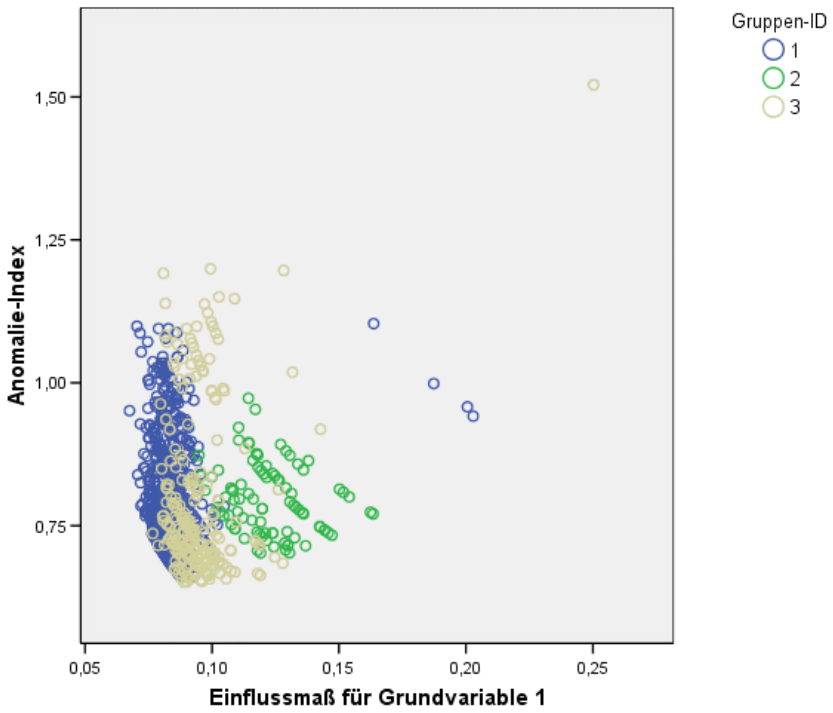
Abbildung 9-16
Dialogfeld "Diagrammerstellung"



- Wählen Sie die Galerie Streu-/Punktdiagramm aus und ziehen Sie das Symbol für gruppierte Streudiagramme auf die Zeichenfläche.
- Wählen Sie *Anomalie-Index* als *y*-Variable und *Einflussmaß für Grundvariable 1* als *x*-Variable aus.
- Wählen Sie *Gruppen-ID* als Variable aus, nach der die Farben gesetzt werden sollen.
- Klicken Sie auf OK.

Nun wird das Streudiagramm erstellt.

Abbildung 9-17
Streudiagramm des Anomalie-Index über das Einflussmaß der ersten Grundvariablen



Das Diagramm ergibt Folgendes:

- Der Fall in der oberen rechten Ecke gehört zu Gruppe 3. Er ist der ungewöhnlichste Fall und zudem der Fall, bei dem eine einzelne Variable den größten Einfluss aufweist.
- Entlang der *y*-Achse ist ersichtlich, dass Gruppe 3 drei Fälle enthält, deren Werte für den Anomalie-Index knapp über 2,00 liegen. Diese Fälle sind potenziell anomal und sollten näher untersucht werden.
- Entlang der *x*-Achse ist ersichtlich, dass Gruppe 1 vier Fälle enthält, deren Variablen-Einflussmaße im Bereich von 0,23 bis 0,33 liegen. Diese Fälle sollten näher untersucht werden, weil diese Werte dazu führen, dass sich die entsprechenden Fälle von den anderen Fällen absetzen.
- Gruppe 2 scheint homogen zu sein: Ihr Anomalie-Index und ihre Variablen-Einflussmaße weichen nicht sehr stark von der zentrale Tendenz ab.

Übersicht

Mit der Prozedur “Ungewöhnliche Fälle identifizieren” haben Sie verschiedene Fälle ausgesondert, die näher untersucht werden sollten. Diese Fälle können mit keinem anderen Validierungsverfahren erkannt werden, weil die Einstufung als anomal nicht nur auf der Grundlage der Variablenwerte, sondern anhand der Beziehungen zwischen den Variablen erfolgt.

Es ist ein wenig enttäuschend, dass die Gruppen weitestgehend auf der Grundlage von zwei Variablen gebildet werden: *Dead on arrival* und *Died in hospital*. In einer weiterführenden Analyse könnten Sie untersuchen, welche Auswirkungen es hat, wenn Sie eine größere Anzahl von Gruppen erzwingen, oder Sie könnten eine Analyse durchführen, die nur auf den überlebenden Patienten beruht.

Verwandte Prozeduren

Die Prozedur “Ungewöhnliche Fälle identifizieren” ist nützlich, um anomale Fälle in einer Datendatei aufzudecken.

- Mit der Prozedur Daten validieren können verdächtige und ungültige Fälle, Variablen und Datenwerte in der Arbeitsdatei identifiziert werden.

Kapitel
10

Optimales Klassieren

Die Prozedur “Optimales Klassieren” diskretisiert eine oder mehrere metrische Variablen (als **Klassierungs-Eingabevariablen** bezeichnet), indem die Werte der einzelnen Variablen auf verschiedene Klassen verteilt werden. Die Klassenbildung ist in Bezug auf eine kategoriale Führungsvariable optimal, die den Klassierungsvorgang “überwacht”. Bei Prozeduren, bei denen kategoriale Variablen erforderlich oder vorzuziehen sind, können dann anstatt der ursprünglichen Datenwerte die Klassen zur weiteren Analyse verwendet werden.

Der Algorithmus für optimales Klassieren

Die Grundschritte für den Algorithmus für optimales Klassieren lassen sich wie folgt charakterisieren:

Vorverarbeitung (optional) Die Klassierungs-Eingabevariable wird in *n* Klassen unterteilt (den Wert für *n* geben Sie selbst an), wobei jede Klasse gleich viele Fälle enthält (bzw. annähernd gleich viele Fälle, wenn sich die Anzahl der Fälle nicht restlos durch *n* teilen lässt).

Ermitteln potenzieller Trennwerte. Jeder unterschiedliche Wert der Klassierungs-Eingabe, der nicht zur selben Kategorie der Führungsvariablen gehört wie der nächstgrößere Wert der Klassierungs-Eingabevariablen, ist ein potenzieller Trennwert.

Auswählen von Trennwerten. Der potenzielle Trennwert, der zum größten Informationsgewinn führt, wird durch das MDLP-Akzeptanzkriterium ausgewertet. Wiederholen Sie den Vorgang, bis keine weiteren potenziellen Trennwerte akzeptiert werden. Die akzeptierten Trennwerte legen die Klassengrenzen fest.

Verwenden der optimalen Klassierung zur Diskretisierung der Daten zu Kreditantragstellern

Im Rahmen der Bemühungen einer Bank, den Anteil der nicht zurückgezahlten Kredite zu reduzieren, hat ein Kreditsachbearbeiter finanzielle und demografische Informationen zu früheren und gegenwärtigen Kunden gesammelt, in der Hoffnung, ein Modell erstellen zu können, das die Wahrscheinlichkeit der Nichtrückzahlung bei Krediten vorhersagt. Mehrere potenzielle Einflussvariablen sind metrisch, der Kreditsachbearbeiter möchte jedoch in der Lage sein, Modelle zu betrachten, die am besten für kategoriale Einflussvariablen geeignet sind.

Informationen zu 5000 früheren Kunden finden Sie in der Datei *bankloan_binning.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 139. Erstellen Sie mithilfe der Prozedur “Optimales Klassieren” Klassierungsregeln für die metrischen Einflussvariablen

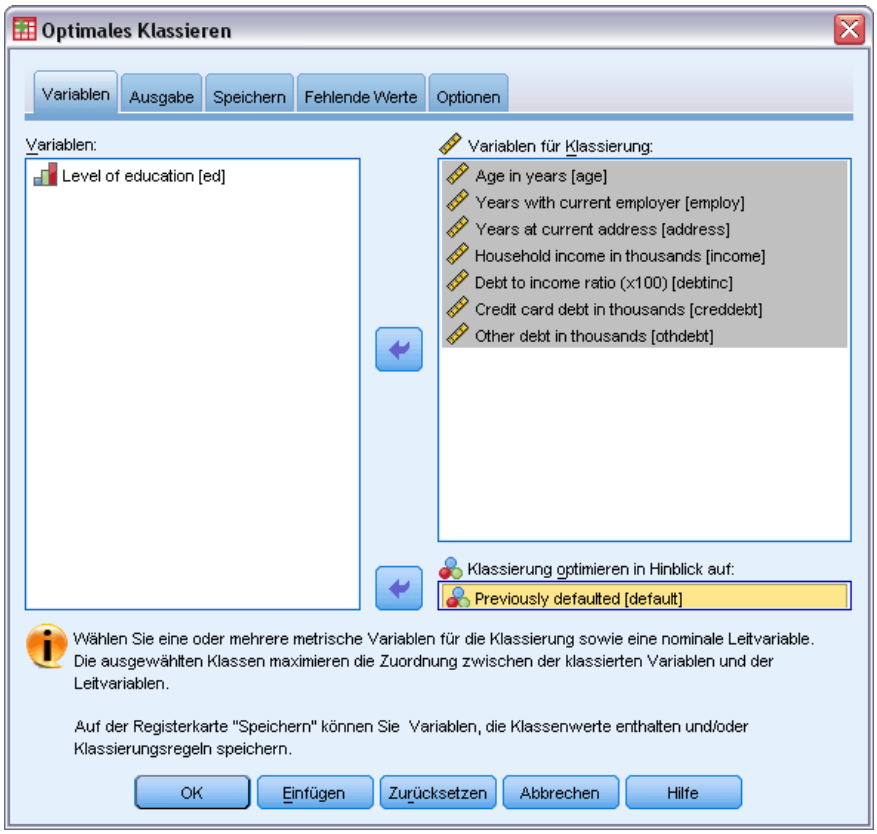
und verwenden Sie diese Regeln anschließend zur Verarbeitung von *bankloan.sav*. Mithilfe des verarbeiteten Daten-Sets kann dann ein Vorhersagemodell erstellt werden.

Durchführen der Analyse

- Zum Ausführen einer Analyse vom Typ “Optimales Klassieren” wählen Sie die folgenden Menübefehle aus:

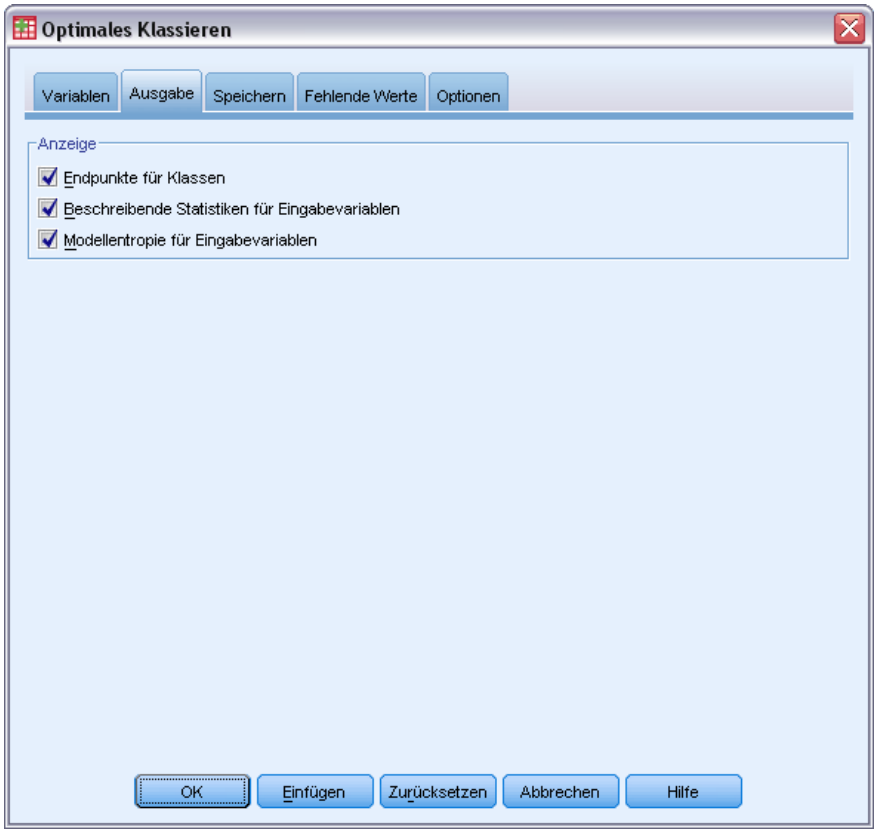
  Transformieren
  Optimales Klassieren...

Abbildung 10-1
Dialogfeld “Optimales Klassieren”, Registerkarte “Variablen”


Optimales Klassieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Variablen:
Level of education [ed]
Variablen für Klassierung:
Age in years [age]
Years with current employer [employ]
Years at current address [address]
Household income in thousands [income]
Debt to income ratio (x100) [debtinc]
Credit card debt in thousands [creddebt]
Other debt in thousands [othdebt]
Klassierung optimieren in Hinblick auf:
Previously defaulted [default]
Wählen Sie eine oder mehrere metrische Variablen für die Klassierung sowie eine nominale Leitvariable. Die ausgewählten Klassen maximieren die Zuordnung zwischen der klassierten Variablen und der Leitvariablen.
Auf der Registerkarte "Speichern" können Sie Variablen, die Klassenwerte enthalten und/oder Klassierungsregeln speichern.
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

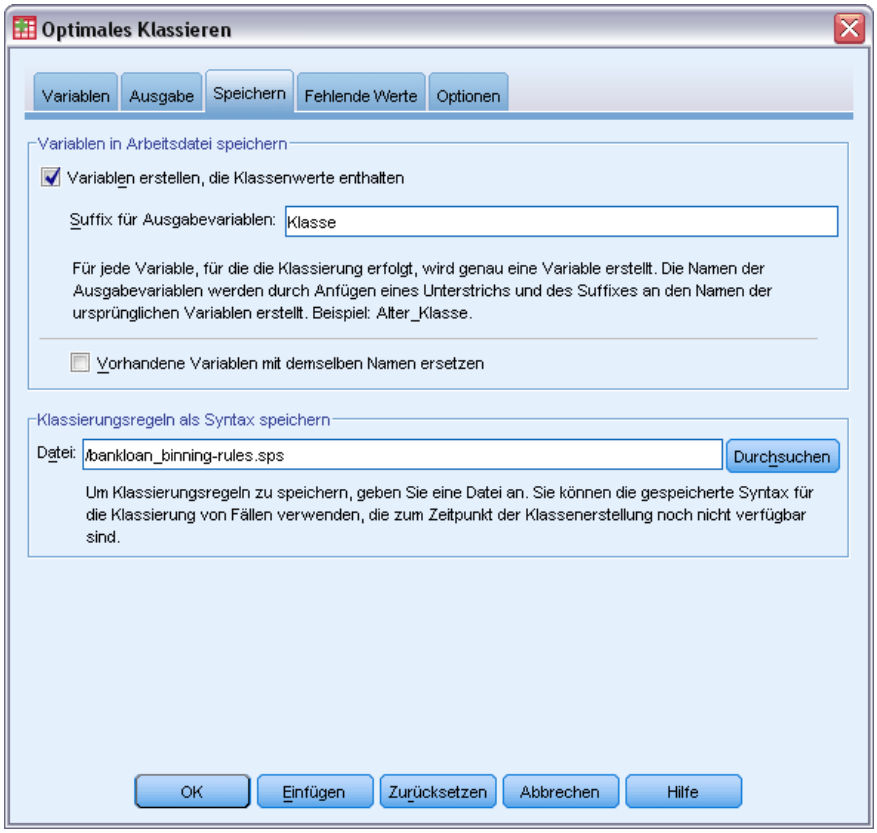
- Wählen Sie *Age in years* (Alter in Jahren) und *Years with current employer* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber) bis *Other debt in thousands* (Andere Schulden in Tausend) als Variablen für die Klassierung aus.
- Wählen Sie *Previously defaulted* (Vorherige Nichtzahlung) als Führungsvariable aus.
- Klicken Sie auf die Registerkarte Ausgabe.

Abbildung 10-2
Dialogfeld "Optimales Klassieren", Registerkarte "Ausgabe"



- Wählen Sie Beschreibende Statistiken und Modellentropie für die zu klassierenden Variablen aus.
- Klicken Sie auf die Registerkarte Speichern.

Abbildung 10-3
Dialogfeld "Optimales Klassieren", Registerkarte "Speichern"


Optimales Klassieren
Variablen
Ausgabe
Speichern
Fehlende Werte
Optionen
Variablen in Arbeitsdatei speichern
Variablen erstellen, die Klassenwerte enthalten
Suffix für Ausgabevariablen:
Klasse
Für jede Variable, für die die Klassierung erfolgt, wird genau eine Variable erstellt. Die Namen der Ausgabevariablen werden durch Anfügen eines Unterstrichs und des Suffixes an den Namen der ursprünglichen Variablen erstellt. Beispiel: Alter_Klasse.
Vorhandene Variablen mit demselben Namen ersetzen
Klassierungsregeln als Syntax speichern
Datei:
/bankloan_binning-rules.sps
Durchsuchen
Um Klassierungsregeln zu speichern, geben Sie eine Datei an. Sie können die gespeicherte Syntax für die Klassierung von Fällen verwenden, die zum Zeitpunkt der Klassenerstellung noch nicht verfügbar sind.
OK
Einfügen
Zurücksetzen
Abbrechen
Hilfe

- Wählen Sie Variablen erstellen, die Werte der Daten in Klassen enthalten.
- Geben Sie einen Pfad und einen Dateinamen für die Syntaxdatei ein, die die generierten Klassierungsregeln enthalten soll. In diesem Beispiel haben wir */bankloan_binning-rules.sps* verwendet.
- Klicken Sie auf OK.

Diese Auswahl führt zu folgender Befehlssyntax:

```
* Optimales Klassieren.
OPTIMAL BINNING
  /VARIABLES GUIDE=default BIN=age employ address income debtinc creddebt
  othdebt SAVE=YES (INTO=age_bin employ_bin address_bin income_bin debtinc_bin
  creddebt_bin othdebt_bin)
  /CRITERIA METHOD=MDLP
   PREPROCESS=EQUALFREQ (BINS=1000)
   FORCEMERGE=0
   LOWERLIMIT=INCLUSIVE
   LOWEREND=UNBOUNDED
   UPPEREND=UNBOUNDED
  /MISSING SCOPE=PAIRWISE
  /OUTFILE RULES='/bankloan_binning-rules.sps'
  /PRINT ENDPOINTS DESCRIPTIVES ENTROPY.
```

- Durch die Prozedur werden die Klassierungs-Eingabevariablen *age*, *employ*, *address*, *income*, *debtinc*, *creddebt* und *othdebt* mithilfe der MDLP-Klassierung mit der Führungsvariablen *default* diskretisiert.
- Die diskretisierten Werte für diese Variablen werden in den neuen Variablen *age_Klasse*, *employ_Klasse*, *address_Klasse*, *income_Klasse*, *debtinc_Klasse*, *creddebt_Klasse* und *othdebt_Klasse* gespeichert.
- Wenn eine Binning-Eingabevariable mehr als 1000 verschiedene Werte aufweist, wird die Anzahl vor der Durchführung der MDLP-KLassierung mithilfe der Methode der gleichen Häufigkeiten auf 1000 reduziert.
- Die Befehlssyntax für die Klassierungsregeln wird in der Datei *c:\bankloan_binning-rules.sps* gespeichert.
- Für die Klassierungs-Eingabevariablen werden die Klassengrenzen und die Modellentropiewerte angefordert.
- Für die anderen Klassierungskriterien werden die Standardwerte verwendet.

Deskriptive Statistiken

Abbildung 10-4
Deskriptive Statistik

	N	Minimum	Maximum	Anzahl der verschiedenen Werte	Anzahl der Klassen
Age in years	5000	20	58	39	2
Years with current employer	5000	0	38	39	4
Years at current address	5000	0	37	38	3
Household income in thousands	5000	12.10	2461.70	1100	2
Debt to income ratio (x100)	5000	.08	44.62	2060	5
Credit card debt in thousands	5000	.01	139.58	5000	4
Other debt in thousands	5000	.01	416.52	4999	2

Die Tabelle “Deskriptive Statistiken” enthält zusammenfassende Informationen zu den Klassierungs-Eingabevariablen. Die ersten vier Spalten betreffen die vorklassierten Werte.

- N ist die Anzahl der in der Analyse verwendeten Fälle. Wenn listenweises Löschen fehlender Werte verwendet wird, sollte dieser Wert für alle Variablen konstant sein. Wenn paarweises Löschen fehlender Werte verwendet wird, ist dieser Wert möglicherweise nicht konstant. Da das vorliegende Daten-Set keine fehlenden Werte aufweist, handelt es sich bei diesem Wert einfach um die Anzahl der Fälle.
- Die Spalten Minimum und Maximum zeigen die Mindest- und Höchstwerte (für Vorklassierung) im Daten-Set für die einzelnen Klassierungs-Eingabevariablen. Durch diese Spalten erhalten Sie nicht nur einen Eindruck von dem beobachteten Wertebereich für die einzelnen Variablen, sondern sie können auch hilfreich beim Aufspüren von Werten sein, die außerhalb des erwarteten Bereichs liegen.
- In der Spalte Anzahl der verschiedenen Werte erfahren Sie, welche Variablen mithilfe des Algorithmus für gleiche Häufigkeiten vorverarbeitet wurden. Standardmäßig werden Variablen mit mehr als 1000 verschiedenen Werten (*Household income in thousands* (Haushaltseinkommen in Tausend) bis *Other debt in thousands* (Andere Schulden in Tausend)) durch die Vorklassierung in 1000 verschiedene Klassen eingeteilt. Diese

vorverarbeiteten Klassen werden anschließend unter Verwendung von MDLP anhand der Führungsvariablen klassiert. Auf der Registerkarte “Optionen” können Sie Einfluss auf die Vorverarbeitungsfunktion nehmen.

- Die Spalte Anzahl der Klassen enthält die endgültige Anzahl an Klassen, die von der Prozedur erstellt werden. Diese ist erheblich kleiner als die Anzahl der verschiedenen Werte.

Modellentropie

Abbildung 10-5
Modellentropie

	Modellentropie
Age in years	.788
Years with current employer	.754
Years at current address	.781
Household income in thousands	.803
Debt to income ratio (x100)	.711
Credit card debt in thousands	.776
Other debt in thousands	.801

Smaller model entropy indicates higher predictive accuracy of the binned variable on guide variable Previously defaulted.

Anhand der Tabelle “Modellentropie” erhalten Sie eine Vorstellung davon, wie nützlich die einzelnen Variablen in einem Vorhersagemodell für die Wahrscheinlichkeit der Nichtzurückzahlung sein könnten.

- Die bestmögliche Einflussvariable ist eine, die für jede generierte Klasse Fälle mit denselben Werten enthält, wie die Führungsvariable, sodass die Führungsvariable perfekt vorhergesagt werden kann. Für eine solche Einflussvariable ist die Modellentropie nicht definiert. Dieser Fall kommt im realen Leben nicht vor und kann auf Probleme mit der Qualität der Daten hindeuten.
- Die schlechtestmögliche Einflussvariable ist eine Variable, deren Verwendung zu keinem besseren Ergebnis führt als bloßes Raten. Der Wert ihrer Modellentropie hängt von den Daten ab. In diesem Datensatz kam es bei 1256 (bzw. 0,2512) der 5000 Kunden zu Schwierigkeiten bei der Kreditrückzahlung, während 3744 (bzw. 0,7488) ihren Kredit zurückzahlten. Die schlechtestmögliche Einflussvariable hätte also eine Modellentropie von $-0{,}2512 \times \log_2(0{,}2512) - 0{,}7488 \times \log_2(0{,}7488) = 0{,}8132$.

Es lässt sich schwerlich eine schlüssigere Aussage treffen, als dass Variablen mit niedrigeren Werten für die Modellentropie besser als Einflussvariablen geeignet sein dürften, da es von der jeweiligen Anwendung und den jeweiligen Daten abhängt, was ein guter Wert für die Modellentropie ist. In diesem Fall haben anscheinend Variablen, die in Bezug auf die Anzahl der unterschiedlichen Kategorien eine größere Anzahl an generierten Klassen aufweisen, niedrigere Werte bei der Modellentropie. Es sollte eine weitere Auswertung dieser Klassierungs-Eingabevariablen als Einflussvariablen durchgeführt werden. Hierfür sollten Prozeduren für Vorhersagemodelle verwendet werden, bei denen eine größere Palette an Werkzeugen für die Variablenauswahl zur Verfügung steht.

Klassierungs-Zusammenfassungen

Die Klassierungs-Zusammenfassung gibt die Grenzen der generierten Klassen und die Häufigkeitszählung für die einzelnen Klassen anhand der Werte der Führungsvariablen wieder. Für jede Klassierungs-Eingabevariable wird eine gesonderte Tabelle mit der Klassierungs-Zusammenfassung erstellt.

Abbildung 10-6
Klassierungs—Zusammenfassung für "Age in Years" (Alter in Jahren)

	Endpunkt		Anzahl der Fälle nach Nivau von Previously defaulted		
Klasse	Minimum	Maximum	No	Yes	Gesamt
1	[a]	32	1129	639	1768
2	32	[a]	2615	617	3232
Gesamt			3744	1256	5000

Jede Klasse wird wie folgt berechnet: Minimum <= Age in years < Maximum.

a. Unbegrenzt

Die Zusammenfassung für *Age in years* (Alter in Jahren) zeigt, dass 1768 Kunden, alle im Alter von 32 Jahren oder darunter, in Klasse 1 eingeteilt wurden, während die übrigen 3232 Kunden, deren Alter jeweils mehr als 32 Jahre beträgt, alle in Klasse 2 eingeteilt wurden. Der Anteil der Kunden, die schon einmal einen Kredit nicht zurückgezahlt haben ("Previously defaulted") ist in Klasse 1 wesentlich höher (639/1768=0,361) als in Klasse 2 (617/3232=0,191).

Abbildung 10-7
Klassierungs-Zusammenfassung für "Household income in thousands" (Haushaltseinkommen in Tausend)

	Endpunkt		Anzahl der Fälle nach Nivau von Previously defaulted		
Klasse	Minimum	Maximum	No	Yes	Gesamt
1	[a]	26,70	1054	513	1567
2	26,70	[a]	2690	743	3433
Gesamt			3744	1256	5000

Jede Klasse wird wie folgt berechnet: Minimum <= Household income in thousands < Maximum.

a. Unbegrenzt

Die Zusammenfassung für *Household income in thousands* (Haushaltseinkommen in Tausend) zeigt ein ähnliches Muster, mit einem einzigen Trennwert bei 26,70 und einem höheren Anteil an Kunden mit früheren Zahlungsschwierigkeiten ("Previously defaulted") in Klasse 1 (513/1567=0,327) als in Klasse 2 (743/3433=0,216). Wie aus der Statistik für die Modellentropie zu erwarten, ist der Unterschied in diesen Anteilen nicht so groß wie bei *Age in years* (Alter in Jahren).

Abbildung 10-8
Klassierungs-Zusammenfassung für "Other debt in thousands" (Andere Schulden in Tausend)

	Endpunkt		Anzahl der Fälle nach Nivau von Previously defaulted		
Klasse	Minimum	Maximum	No	Yes	Gesamt
1	[a]	2,19	2161	539	2700
2	2,19	[a]	1583	717	2300
Gesamt			3744	1256	5000

Jede Klasse wird wie folgt berechnet: Minimum <= Other debt in thousands < Maximum.

a. Unbegrenzt

Die Zusammenfassung für *Other debt in thousands* (Andere Schulden in Tausend) zeigt ein umgekehrtes Muster, mit einem einzigen Trennwert bei 2,19 und einem geringeren Anteil an Kunden mit früheren Zahlungsschwierigkeiten ("Previously defaulted") in Klasse 1 (539/2700=0,200) als in Klasse 2 (717/2300=0,312). Auch hier ist, wie aus der Statistik für die Modellentropie zu erwarten, der Unterschied in diesen Anteilen nicht so groß wie bei *Age in years* (Alter in Jahren).

Abbildung 10-9
Klassierungs-Zusammenfassung für "Years with current employer" (Jahre der Beschäftigung beim derzeitigen Arbeitgeber)

	Endpunkt		Anzahl der Fälle nach Nivau von Previously defaulted		
Klasse	Minimum	Maximum	No	Yes	Gesamt
1	[a]	3	629	478	1107
2	3	8	1066	461	1527
3	8	18	1471	268	1739
4	18	[a]	578	49	627
Gesamt			3744	1256	5000

Jede Klasse wird wie folgt berechnet: Minimum <= Years with current employer < Maximum.

a. Unbegrenzt

Die Zusammenfassung für *Years with current employer* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber) zeigt ein Muster abnehmender Anteile der zahlungsunfähigen Personen bei steigender Klassenzahl.

Klasse	**Anteil der zahlungsunfähigen Personen**
1	0.432
2	0.302
3	0.154
4	0.078

Abbildung 10-10
Klassierungs-Zusammenfassung für "Years at current address" (Wohnhaft an gleicher Adresse (in Jahren))

	Endpunkt		Anzahl der Fälle nach Nivau von Previously defaulted		
Klasse	Minimum	Maximum	No	Yes	Gesamt
1	[a]	7	1652	829	2481
2	7	14	1184	313	1497
3	14	[a]	908	114	1022
Gesamt			3744	1256	5000

Jede Klasse wird wie folgt berechnet: Minimum <= Years at current address < Maximum.

a. Unbegrenzt

Die Zusammenfassung für *Years at current address* (Wohnhaft an gleicher Adresse (in Jahren)) zeigt ein ähnliches Muster. Wie aus der Statistik für die Modellentropie zu erwarten, sind die Unterschiede zwischen den Klassen beim Anteil der zahlungsunfähigen Personen bei *Years with current employer* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber) deutlicher als bei *Years at current address* (Wohnhaft an gleicher Adresse (in Jahren)).

Klasse	**Anteil der zahlungsunfähigen Personen**
1	0.334
2	0.209
3	0.112

Abbildung 10-11
Klassierungs-Zusammenfassung für "Credit card debt in thousands" (Schulden auf Kreditkarte in Tausend)

	Endpunkt		Anzahl der Fälle nach Nivau von Previously defaulted		
Klasse	Minimum	Maximum	No	Yes	Gesamt
1	[a]	,97	2169	466	2635
2	,97	1,91	848	307	1155
3	1,91	6,05	643	352	995
4	6,05	[a]	84	131	215
Gesamt			3744	1256	5000

Jede Klasse wird wie folgt berechnet: Minimum <= Credit card dept in thousands < Maximum.

a. Unbegrenzt

Die Zusammenfassung für *Credit card debt in thousands* (Schulden auf Kreditkarte in Tausend) zeigt das umgekehrte Muster: bei steigender Klassenzahl nehmen die Anteile der zahlungsunfähigen Personen zu. Die Variablen *Years with current employer* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber) und *Years at current address* (Wohnhaft an gleicher Adresse (in Jahren)) scheinen besser zur Ermittlung von Personen geeignet, die mit großer Wahrscheinlichkeit nicht in Zahlungsschwierigkeiten geraten, während *Credit card debt in*

thousands (Schulden auf Kreditkarte in Tausend) besser für die Ermittlung von Personen geeignet ist, die mit großer Wahrscheinlichkeit den Kredit nicht zurückzahlen können.

Klasse	**Anteil der zahlungsunfähigen Personen**
1	0.177
2	0.266
3	0.354
4	0.609

Abbildung 10-12
Klassierungs-Zusammenfassung für "Debt to income ratio (x100)" (Relation Schulden zu Einkommen (in %))

	Endpunkt		Anzahl der Fälle nach Niveau von Previously defaulted		
Klasse	Minimum	Maximum	No	Yes	Gesamt
1	[a]	4,39	912	88	1000
2	4,39	12,09	2006	437	2443
3	12,09	18,71	625	386	1011
4	18,71	31,00	198	303	501
5	31,00	[a]	3	42	45
Gesamt			3744	1256	5000

Jede Klasse wird wie folgt berechnet: Minimum <= Debt to income ratio (x100) < Maximum.

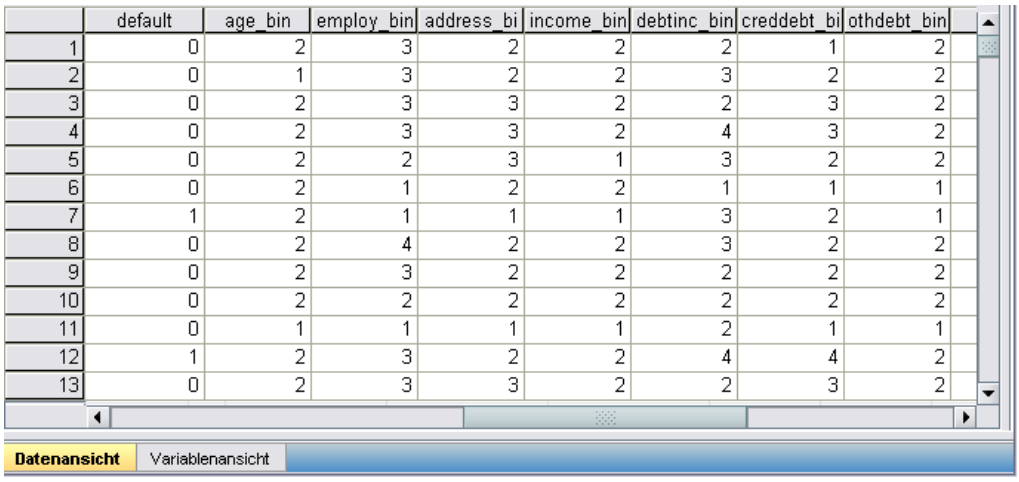
a. Unbegrenzt

Die Zusammenfassung für *Debt to income ratio (x100)* (Relation Schulden zu Einkommen (in %)) weist ein ähnliches Muster auf wie *Credit card debt in thousands* (Schulden auf Kreditkarte in Tausend). Diese Variable weist den niedrigsten Wert für die Modellentropie auf und ist somit der beste Kandidat als Einflussvariable für die Wahrscheinlichkeit der Zahlungsunfähigkeit. Sie bietet eine bessere Klassifizierung von Personen, die mit großer Wahrscheinlichkeit zahlungsunfähig werden, als *Credit card debt in thousands* (Schulden auf Kreditkarte in Tausend) und eine fast ebenso gute Klassifizierung von Personen, die mit geringer Wahrscheinlichkeit zahlungsunfähig werden, wie *Years with current employer* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber).

Klasse	**Anteil der zahlungsunfähigen Personen**
1	0.088
2	0.179
3	0.382
4	0.605
5	0.933

Klassierte Variablen

Abbildung 10-13
Klassierte Variablen für bankloan_binning.sav im Daten-Editor



	default	age_bin	employ_bin	address_bi	income_bin	debtinc_bin	creddebt_bi	othdebt_bin
1	0	2	3	2	2	2	1	2
2	0	1	3	2	2	3	2	2
3	0	2	3	3	2	2	3	2
4	0	2	3	3	2	4	3	2
5	0	2	2	3	1	3	2	2
6	0	2	1	2	2	1	1	1
7	1	2	1	1	1	3	2	1
8	0	2	4	2	2	3	2	2
9	0	2	3	2	2	2	2	2
10	0	2	2	2	2	2	2	2
11	0	1	1	1	1	2	1	1
12	1	2	3	2	2	4	4	2
13	0	2	3	3	2	2	3	2

Datenansicht | Variablenansicht

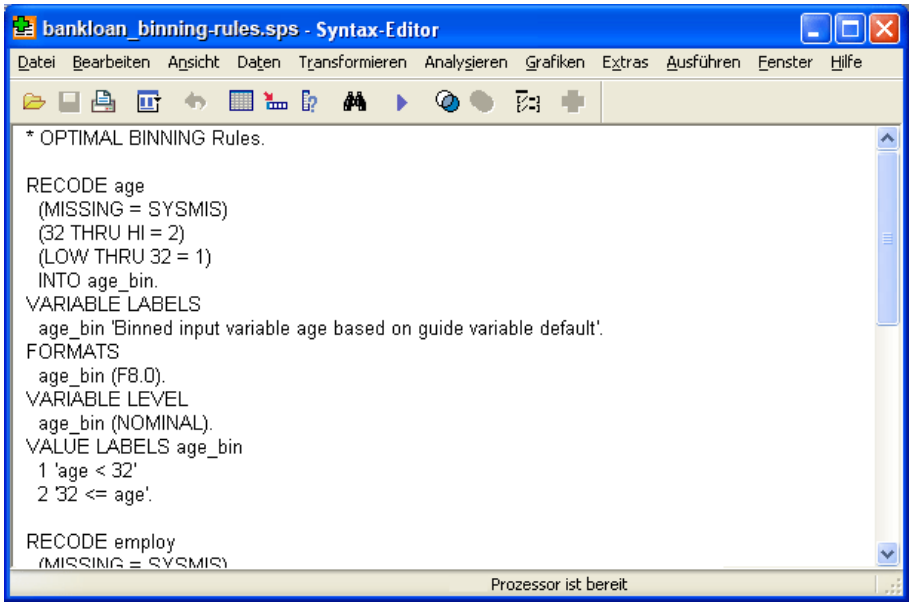
Die Ergebnisse des Klassierungsprozesses für dieses Daten-Set werden aus dem Daten-Editor ersichtlich. Diese klassierten Variablen sind nützlich, wenn Sie benutzerdefinierte Zusammenfassungen der Klassierungsergebnisse mithilfe von deskriptiven Prozeduren oder Berichtsprozeduren erstellen möchten. Es ist jedoch nicht ratsam, dieses Daten-Set zur Erstellung eines Vorhersagemodells zu verwenden, da die Klassierungsregeln mithilfe dieser Fälle erstellt wurden. Es ist sinnvoller, die Klassierungsregeln auf ein anderes Daten-Set anzuwenden, das Informationen zu anderen Kunden enthält.

Anwenden von Syntax-Klassierungsregeln

Bei der Ausführung der Prozedur “Optimales Klassieren” haben Sie angegeben, dass die von der Prozedur erstellten Klassierungsregeln als Befehlssyntax gespeichert werden sollten.

► Öffnen Sie die Datei *bankloan_binning-rules.sps*.

Abbildung 10-14
Syntaxregeldatei



Für jede Klassierungs-Eingabevariable gibt es einen Block mit Befehlssyntax, die die Klassierung durchführt, Variablenlabel, Format und Stufe und die Variablenlabels für die Klassen festlegt. Diese Befehle können auf ein Daten-Set angewendet werden, das dieselben Variablen enthält wie *bankloan_binning.sav*.

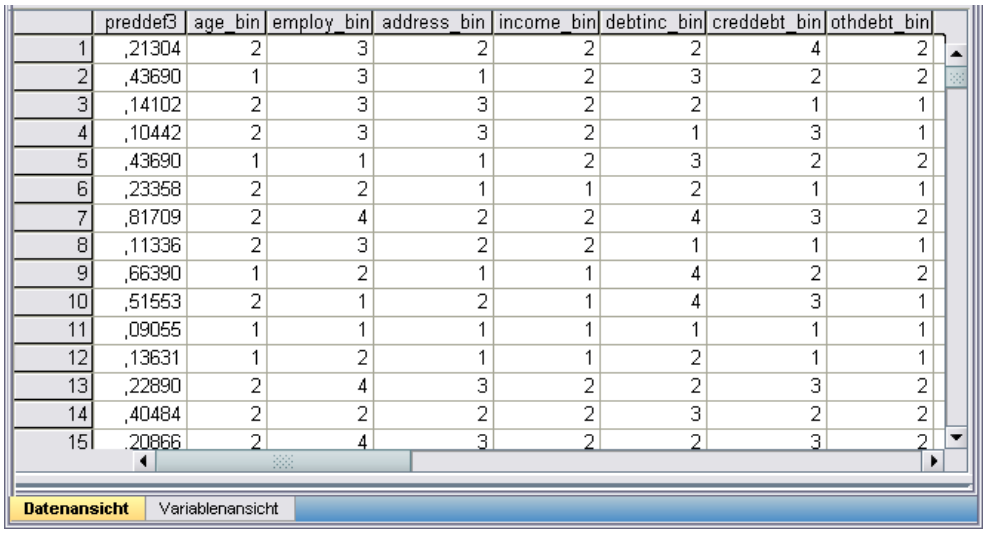
- Öffnen Sie die Datei *bankloan.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 139.

- Kehren Sie zur Syntax-Editor-Ansicht von *bankloan_binning-rules.sps* zurück.

▶ Um die Klassierungsregeln anzuwenden, wählen Sie im Syntax-Editor folgende Befehle aus:

```
Ausführen
  Alles...
```

Abbildung 10-15
Klassierte Variablen für bankloan.sav im Daten-Editor



	preddef3	age_bin	employ_bin	address_bin	income_bin	debtinc_bin	creddebt_bin	othdebt_bin
1	,21304	2	3	2	2	2	4	2
2	,43690	1	3	1	2	3	2	2
3	,14102	2	3	3	2	2	1	1
4	,10442	2	3	3	2	1	3	1
5	,43690	1	1	1	2	3	2	2
6	,23358	2	2	1	1	2	1	1
7	,81709	2	4	2	2	4	3	2
8	,11336	2	3	2	2	1	1	1
9	,66390	1	2	1	1	4	2	2
10	,51553	2	1	2	1	4	3	1
11	,09055	1	1	1	1	1	1	1
12	,13631	1	2	1	1	2	1	1
13	,22890	2	4	3	2	2	3	2
14	,40484	2	2	2	2	3	2	2
15	,20866	2	4	3	2	2	3	2

Datenansicht | Variablenansicht

Die Variablen in*bankloan.sav* wurden klassiert. Hierfür wurden die Regeln verwendet, die bei der Ausführung der Prozedur “Optimales Klassieren” für die Datei *bankloan_binning.sav* erstellt wurden. Dieses Daten-Set kann nun zur Erstellung von Vorhersagemodellen verwendet werden, bei denen kategoriale Variablen erforderlich oder vorzuziehen sind.

Zusammenfassung

Mithilfe der Prozedur “Optimales Klassieren” haben wir Klassierungsregeln für metrische Variablen generiert, die potenzielle Einflussvariablen für die Wahrscheinlichkeit der Zahlungsunfähigkeit sind, und haben diese Regeln auf ein separates Daten-Set angewendet.

Während des Klassierungsvorgangs haben wir festgestellt, dass die klassierten Variablen *Years with current employer* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber) und *Years at current address* (Wohnhaft an gleicher Adresse (in Jahren)) besser zur Ermittlung von Personen geeignet sind, die mit großer Wahrscheinlichkeit nicht in Zahlungsschwierigkeiten geraten, während *Credit card debt in thousands* (Schulden auf Kreditkarte in Tausend) besser für die Ermittlung von Personen geeignet ist, die mit großer Wahrscheinlichkeit den Kredit nicht zurückzahlen können. Diese interessante Beobachtung ist sehr wertvoll, wenn es darum geht, Vorhersagemodelle für die Wahrscheinlichkeit der Zahlungsunfähigkeit zu erstellen. Wenn die Vermeidung uneinbringlicher Forderungen das Hauptanliegen ist, ist die Variable *Credit card debt in thousands* (Schulden auf Kreditkarte in Tausend) wichtiger als *Years with current employer* (Jahre der Beschäftigung beim derzeitigen Arbeitgeber) und *Years at current address* (Wohnhaft an gleicher Adresse (in Jahren)) . Wenn die Erweiterung des Kundenstamms oberste Priorität hat, sind die Variablen *Years with current employer*(Jahre der Beschäftigung beim derzeitigen Arbeitgeber) und *Years at current address* (Wohnhaft an gleicher Adresse (in Jahren)) von größerer Bedeutung.

Anhang

A

Beispieldateien

Die zusammen mit dem Produkt installierten Beispieldateien finden Sie im Unterverzeichnis *Samples* des Installationsverzeichnisses. Für jeder der folgenden Sprachen gibt es einen eigenen Ordner innerhalb des Unterverzeichnisses "Samples": Englisch, Französisch, Deutsch, Italienisch, Japanisch, Koreanisch, Polnisch, Russisch, Vereinfachtes Chinesisch, Spanisch und Traditionelles Chinesisch.

Nicht alle Beispieldateien stehen in allen Sprachen zur Verfügung. Wenn eine Beispieldatei nicht in einer Sprache zur Verfügung steht, enthält der jeweilige Sprachordner eine englische Version der Beispieldatei.

Beschreibungen

Im Folgenden finden Sie Kurzbeschreibungen der in den verschiedenen Beispielen in der Dokumentation verwendeten Beispieldateien.

- **accidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die alters- und geschlechtsabhängige Risikofaktoren für Autounfälle in einer bestimmten Region untersucht. Jeder Fall entspricht einer Kreuzklassifikation von Alterskategorie und Geschlecht.
- **adl.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die Vorteile einer vorgeschlagenen Therapieform für Schlaganfallpatienten zu ermitteln. Ärzte teilten weibliche Schlaganfallpatienten nach dem Zufallsprinzip jeweils einer von zwei Gruppen zu. Die erste Gruppe erhielt die physische Standardtherapie, die zweite erhielt eine zusätzliche Emotionaltherapie. Drei Monate nach den Behandlungen wurden die Fähigkeiten der einzelnen Patienten, übliche Alltagsaktivitäten auszuführen, als ordinale Variablen bewertet.
- **advert.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Einzelhändlers geht, die Beziehungen zwischen den in Werbung investierten Beträgen und den daraus resultierenden Umsätzen zu untersuchen. Zu diesem Zweck hat er die Umsätze vergangener Jahre und die zugehörigen Werbeausgaben zusammengestellt.
- **aflatoxin.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests von Maisernten auf Aflatoxin geht, ein Gift, dessen Konzentration stark zwischen und innerhalb von Ernteerträgen schwankt. Ein Kornverarbeitungsbetrieb hat aus 8 Ernteerträgen je 16 Proben erhalten und das Aflatoxinniveau in Teilen pro Milliarde (parts per billion, PPB) gemessen.
- **aflatoxin20.sav.** Diese Datendatei enthält die Aflatoxinmessungen aus jeder der 16 Stichproben aus den Erträgen 4 und 8 der Datendatei *aflatoxin.sav*.

- **anorectic.sav.** Bei der Ausarbeitung einer standardisierten Symptomatologie anorektischen/bulimischen Verhaltens führten Forscher (Van der Ham, Meulman, Van Strien, als auch Van Engeland, 1997)) eine Studie mit 55 Jugendlichen mit bekannten Ess-Störungen durch. Jeder Patient wurde vier Mal über einen Zeitraum von vier Jahren untersucht, es fanden also insgesamt 220 Beobachtungen statt. Bei jeder Beobachtung erhielten die Patienten Scores für jedes von 16 Symptomen. Die Symptomwerte fehlen für Patient 71 zum Zeitpunkt 2, Patient 76 zum Zeitpunkt 2 und Patient 47 zum Zeitpunkt 3, wodurch 217 gültige Beobachtungen verbleiben.
- **autoaccidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Versicherungsanalysten geht, ein Modell zur Anzahl der Autounfälle pro Fahrer unter Berücksichtigung von Alter und Geschlecht zu erstellen. Jeder Fall stellt einen Fahrer dar und erfasst das Geschlecht des Fahrers, sein Alter in Jahren und die Anzahl der Autounfälle in den letzten fünf Jahren.
- **band.sav.** Diese Datendatei enthält die hypothetischen wöchentlichen Verkaufszahlen von CDs für eine Musikgruppe. Daten für drei mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **bankloan.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, den Anteil der nicht zurückgezahlten Kredite zu reduzieren. Die Datei enthält Informationen zum Finanzstatus und demografischen Hintergrund von 850 früheren und potenziellen Kunden. Bei den ersten 700 Fällen handelt es sich um Kunden, denen bereits ein Kredit gewährt wurde. Bei den letzten 150 Fällen handelt es sich um potenzielle Kunden, deren Kreditrisiko die Bank als gering oder hoch einstufen möchte.
- **bankloan_binning.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen zum Finanzstatus und demografischen Hintergrund von 5.000 früheren Kunden enthält.
- **behavior.sav.** In einem klassischen Beispiel ((Price als auch Bouffard, 1974)) wurden 52 Schüler/Studenten gebeten, die Kombinationen aus 15 Situationen und 15 Verhaltensweisen auf einer 10-Punkte-Skala von 0 = "ausgesprochen angemessen" bis 9 = "ausgesprochen unangemessen" zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Unähnlichkeiten verwendet.
- **behavior_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine zweidimensionale Lösung für *behavior.sav*.
- **brakes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik geht, die Scheibenbremsen für Hochleistungsautomobile herstellt. Die Datendatei enthält Messungen des Durchmessers von 16 Scheiben aus 8 Produktionsmaschinen. Der Zieldurchmesser für die Scheiben ist 322 Millimeter.
- **breakfast.sav.** In einer klassischen Studie ((Green als auch Rao, 1972)) wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = "am meisten bevorzugt" bis 15 = "am wenigsten bevorzugt" zu ordnen. Die Bevorzugungen wurden in sechs unterschiedlichen Szenarien erfasst, von "Overall preference" (Allgemein bevorzugt) bis "Snack, with beverage only" (Imbiss, nur mit Getränk).
- **breakfast-overall.sav.** Diese Datei enthält die Daten zu den bevorzugten Frühstücksartikeln, allerdings nur für das erste Szenario, "Overall preference" (Allgemein bevorzugt).

- **broadband_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die die Anzahl der Abonnenten eines Breitband-Service, nach Region geordnet, enthält. Die Datendatei enthält die monatlichen Abonnentenzahlen für 85 Regionen über einen Zeitraum von vier Jahren.
- **broadband_2.sav.** Diese Datendatei stimmt mit *broadband_1.sav* überein, enthält jedoch Daten für weitere drei Monate.
- **car_insurance_claims.sav.** Ein an anderer Stelle ((McCullagh als auch Nelder, 1989)) vorgestelltes und analysiertes Daten-Set bezieht sich auf Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit Gamma-Verteilung modellieren. Dazu wird eine inverse Verknüpfungsfunktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination aus Alter des Versicherungsnehmers, Fahrzeugtyp und Fahrzeugalter in Bezug zu setzen. Die Anzahl der eingereichten Schadensansprüche kann als Skalierungsgewicht verwendet werden.
- **car_sales.sav.** Diese Datendatei enthält hypothetische Verkaufsschätzer, Listenpreise und physische Spezifikationen für verschiedene Fahrzeugfabrikate und -modelle. Die Listenpreise und physischen Spezifikationen wurden von *edmunds.com* und Hersteller-Websites entnommen.
- **car_sales_uprepared.sav.** Hierbei handelt es sich um eine modifizierte Version der Datei *car_sales.sav*, die keinerlei transformierte Versionen der Felder enthält.
- **carpet.sav.** In einem beliebten Beispiel möchte (Green als auch Wind, 1973) einen neuen Teppichreiniger vermarkten und dazu den Einfluss von fünf Faktoren auf die Bevorzugung durch den Verbraucher untersuchen: Verpackungsgestaltung, Markenname, Preis, Gütesiegel, *Good Housekeeping* und Geld-zurück-Garantie. Die Verpackungsgestaltung setzt sich aus drei Faktorenebenen zusammen, die sich durch die Position der Auftragebürste unterscheiden. Außerdem gibt es drei Markennamen (*K2R*, *Glory* und *Bissell*), drei Preisstufen sowie je zwei Ebenen (Nein oder Ja) für die letzten beiden Faktoren. 10 Kunden stufen 22 Profile ein, die durch diese Faktoren definiert sind. Die Variable *Preference* enthält den Rang der durchschnittlichen Einstufung für die verschiedenen Profile. Ein niedriger Rang bedeutet eine starke Bevorzugung. Diese Variable gibt ein Gesamtmaß der Bevorzugung für die Profile an.
- **carpet_prefs.sav.** Diese Datendatei beruht auf denselben Beispielen, wie für *carpet.sav* beschrieben, enthält jedoch die tatsächlichen Einstufungen durch jeden der 10 Kunden. Die Kunden wurden gebeten, die 22 Produktprofile in der Reihenfolge ihrer Präferenzen einzustufen. Die Variablen *PREF1* bis *PREF22* enthalten die IDs der zugeordneten Profile, wie in *carpet_plan.sav* definiert.
- **catalog.sav.** Diese Datendatei enthält hypothetische monatliche Verkaufszahlen für drei Produkte, die von einem Versandhaus verkauft werden. Daten für fünf mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **catalog_seasfac.sav.** Diese Datendatei ist mit *catalog.sav* identisch, außer, dass ein Set von saisonalen Faktoren, die mithilfe der Prozedur "Saisonale Zerlegung" berechnet wurden, sowie die zugehörigen Datumsvariablen hinzugefügt wurden.
- **cellular.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Mobiltelefonunternehmens geht, die Kundenabwanderung zu verringern. Scores für die Abwanderungsneigung (von 0 bis 100) werden auf die Kunden angewendet. Kunden mit einem Score von 50 oder höher streben vermutlich einen Anbieterwechsel an.

- **ceramics.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Herstellers geht, der ermitteln möchte, ob ein neue, hochwertige Keramiklegierung eine größere Hitzebeständigkeit aufweist als eine Standardlegierung. Jeder Fall entspricht einem Test einer der Legierungen; die Temperatur, bei der das Keramikwälzlager versagte, wurde erfasst.
- **cereal.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Umfrage geht, bei der 880 Personen nach ihren Frühstückgewohnheiten befragt wurden. Außerdem wurden Alter, Geschlecht, Familienstand und Vorliegen bzw. Nichtvorliegen eines aktiven Lebensstils (auf der Grundlage von mindestens zwei Trainingseinheiten pro Woche) erfasst. Jeder Fall entspricht einem Teilnehmer.
- **clothing_defects.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Bekleidungsfabrik geht. Aus jeder in der Fabrik produzierten Charge entnehmen die Kontrolleure eine Stichprobe an Bekleidungsartikeln und zählen die Anzahl der Bekleidungsartikel die inakzeptabel sind.
- **coffee.sav.** Diese Datendatei enthält Daten zum wahrgenommenen Image von sechs Eiskaffeemarken ((Kennedy, Riquier, als auch Sharp, 1996)). Bei den 23 Attributen des Eiskaffee-Image sollten die Teilnehmer jeweils alle Marken auswählen, die durch dieses Attribut beschrieben werden. Die sechs Marken werden als “AA”, “BB”, “CC”, “DD”, “EE” und “FF” bezeichnet, um Vertraulichkeit zu gewährleisten.
- **contacts.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Kontaktlisten einer Gruppe von Vertretern geht, die Computer an Unternehmen verkaufen. Die einzelnen Kontaktpersonen werden anhand der Abteilung, in der sie in ihrem Unternehmen arbeiten und anhand ihrer Stellung in der Unternehmenshierarchie in Kategorien eingeteilt. Außerdem werden der Betrag des letzten Verkaufs, die Zeit seit dem letzten Verkauf und die Größe des Unternehmens, in dem die Kontaktperson arbeitet, aufgezeichnet.
- **creditpromo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Kaufhauses geht, die Wirksamkeit einer kürzlich durchgeführten Kreditkarten-Werbeaktion einzuschätzen. Dazu wurden 500 Karteninhaber nach dem Zufallsprinzip ausgewählt. Die Hälfte erhielt eine Werbebeilage, die einen reduzierten Zinssatz für Einkäufe in den nächsten drei Monaten ankündigte. Die andere Hälfte erhielt eine Standard-Werbebeilage.
- **customer_dbase.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, das die Informationen in seinem Data Warehouse nutzen möchte, um spezielle Angebote für Kunden zu erstellen, die mit der größten Wahrscheinlichkeit darauf ansprechen. Nach dem Zufallsprinzip wurde eine Untergruppe des Kundenstamms ausgewählt. Diese Gruppe erhielt die speziellen Angebote und die Reaktionen wurden aufgezeichnet.
- **customer_information.sav.** Eine hypothetische Datendatei mit Kundenmailingdaten wie Name und Adresse.
- **customers_model.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, auf die sich eine Marketingkampagne richtete. Zu diesen Daten gehören demografische Informationen, eine Übersicht über die bisherigen Einkäufe und die Angabe ob die einzelnen Personen auf die Kampagne ansprachen oder nicht. Jeder Fall entspricht einer Einzelperson.

- **customers_new.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, die potenzielle Kandidaten für Marketingkampagnen sind. Zu diesen Daten gehören demografische Informationen und eine Übersicht über die bisherigen Einkäufe für jede Person. Jeder Fall entspricht einer Einzelperson.
- **debate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die gepaarte Antworten auf eine Umfrage unter den Zuhörern einer politischen Debatte enthält (Antworten vor und nach der Debatte). Jeder Fall entspricht einem Befragten.
- **debate_aggregate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der die Antworten aus *debate.sav* aggregiert wurden. Jeder Fall entspricht einer Kreuzklassifikation der bevorzugten Politiker vor und nach der Debatte.
- **demo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Kundendatenbank geht, die zum Zwecke der Zusendung monatlicher Angebote erworben wurde. Neben verschiedenen demografischen Informationen ist erfasst, ob der Kunde auf das Angebot geantwortet hat.
- **demo_cs_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den ersten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einer anderen Stadt. Außerdem sind IDs für Region, Provinz, Landkreis und Stadt erfasst.
- **demo_cs_2.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den zweiten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einem anderen Stadtteil aus den im ersten Schritt ausgewählten Städten. Außerdem sind IDs für Region, Provinz, Landkreis, Stadt, Stadtteil und Wohneinheit erfasst. Die Informationen zur Stichprobenziehung aus den ersten beiden Stufen des Stichprobenplans sind ebenfalls enthalten.
- **demo_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfrageinformationen enthält die mit einem komplexen Stichprobenplan erfasst wurden. Jeder Fall entspricht einer anderen Wohneinheit. Es sind verschiedene Informationen zum demografischen Hintergrund und zur Stichprobenziehung erfasst.
- **dmdata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen über Demografie und Einkäufe für ein Direktmarketing-Unternehmen enthält.
- **dietstudy.sav.** Diese hypothetische Datendatei enthält die Ergebnisse einer Studie der "Stillman-Diät" (Rickman, Mitchell, Dingman, als auch Dalen, 1974). Jeder Fall entspricht einem Teilnehmer und enthält dessen Gewicht vor und nach der Diät in amerikanischen Pfund sowie mehrere Messungen des Triglyceridspiegels (in mg/100 ml).
- **dischargedata.sav.** Hierbei handelt es sich um eine Datendatei zum Thema *Seasonal Patterns of Winnipeg Hospital Use*, (Menec , Roos, Nowicki, MacWilliam, Finlayson , als auch Schwarz, 1999) (Saisonale Muster der Belegung im Krankenhaus von Winnipeg) vom Manitoba Centre for Health Policy.
- **dvdplayer.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Entwicklung eines neuen DVD-Spielers geht. Mithilfe eines Prototyps hat das Marketing-Team Zielgruppendaten erfasst. Jeder Fall entspricht einem befragten Benutzer und enthält demografische Daten zu dem Benutzer sowie dessen Antworten auf Fragen zum Prototyp.
- **flying.sav.** Diese Datendatei enthält die Flugmeilen zwischen zehn Städten in den USA.

- **german_credit.sav.** Diese Daten sind aus dem Daten-Set "German credit" im Repository of Machine Learning Databases ((Blake als auch Merz, 1998)) an der Universität von Kalifornien in Irvine entnommen.
- **grocery_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *grocery_coupons.sav*, wobei die wöchentlichen Einkäufe zusammengefasst sind, sodass jeder Fall einem anderen Kunden entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und der verzeichnete ausgegebene Betrag ist nun die Summe der Beträge, die in den vier Wochen der Studie ausgegeben wurden.
- **grocery_coupons.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfragedaten enthält, die von einer Lebensmittelkette erfasst wurden, die sich für die Kaufgewohnheiten ihrer Kunden interessiert. Jeder Kunde wird über vier Wochen beobachtet, und jeder Fall entspricht einer Kundenwoche und enthält Informationen zu den Geschäften, in denen der Kunde einkauft sowie zu anderen Merkmalen, beispielsweise welcher Betrag in der betreffenden Woche für Lebensmittel ausgegeben wurde.
- **guttman.sav.** Bell ((Bell, 1961)) legte eine Tabelle zur Darstellung möglicher sozialer Gruppen vor. Guttman ((Guttman, 1968)) verwendete einen Teil dieser Tabelle, bei der fünf Variablen, die Aspekte beschreiben, wie soziale Interaktion, das Gefühl der Gruppenzugehörigkeit, die physische Nähe der Mitglieder und die Formalität der Beziehung, mit sieben theoretischen sozialen Gruppen gekreuzt wurden: "crowds" (Menschenmassen, beispielsweise die Zuschauer eines Fußballspiels), "audience" (Zuhörerschaften, beispielsweise die Personen im Theater oder bei einer Vorlesung), "public" (Öffentlichkeit, beispielsweise Zeitungsleser oder Fernsehzuschauer), "mobs" (Mobs, wie Menschenmassen, jedoch mit wesentlich stärkerer Interaktion), "primary groups" (Primärgruppen, vertraulich), "secondary groups" (Sekundärgruppen, freiwillig) und "modern community" (die moderne Gesellschaft, ein lockerer Zusammenschluss, der aus einer engen physischen Nähe und dem Bedarf an spezialisierten Dienstleistungen entsteht).
- **healthplans.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Versicherungsgruppe geht, vier verschiedene Pläne zur Gesundheitsvorsorge für Kleinbetriebe zu evaluieren. Zwölf Inhaber von Kleinbetrieben (Arbeitgeber) wurden gebeten, die Pläne danach in eine Rangfolge zu bringen, wie gern sie sie ihren Mitarbeitern anbieten würden. Jeder Fall entspricht einem Arbeitgeber und enthält die Reaktionen auf die einzelnen Pläne.
- **health_funding.sav.** Hierbei handelt es sich um eine hypothetische Datei, die Daten zur Finanzierung des Gesundheitswesens (Betrag pro 100 Personen), Krankheitsraten (Rate pro 10.000 Personen der Bevölkerung) und Besuche bei medizinischen Einrichtungen/Ärzten (Rate pro 10.000 Personen der Bevölkerung) enthält. Jeder Fall entspricht einer anderen Stadt.
- **hivassay.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Bemühungen eines pharmazeutischen Labors, einen Schnelltest zur Erkennung von HIV-Infektionen zu entwickeln. Die Ergebnisse des Tests sind acht kräftiger werdende Rotschattierungen, wobei kräftigeren Schattierungen auf eine höhere Infektionswahrscheinlichkeit hindeuten. Bei 2.000 Blutproben, von denen die Hälfte mit HIV infiziert war, wurde ein Labortest durchgeführt.
- **hourlywagedata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zum Stundenlohn von Pflegepersonal in Praxen und Krankenhäusern mit unterschiedlich langer Berufserfahrung.

- **insurance_claims.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die ein Modell zur Kennzeichnung verdächtiger, potenziell betrügerischer Ansprüche erstellen möchte. Jeder Fall entspricht einem Anspruch.
- **insure.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die die Risikofaktoren untersucht, die darauf hinweisen, ob ein Kunde die Leistungen einer mit einer Laufzeit von 10 Jahren abgeschlossenen Lebensversicherung in Anspruch nehmen wird. Jeder Fall in der Datendatei entspricht einem Paar von Verträgen, je einer mit Leistungsforderung und der andere ohne, wobei die beiden Versicherungsnehmer in Alter und Geschlecht übereinstimmen.
- **judges.sav.** Hierbei handelt es sich um eine hypothetische Datendatei mit den Wertungen von ausgebildeten Kampfrichtern (sowie eines Sportliebhabers) zu 300 Kunstturnleistungen. Jede Zeile stellt eine Leistung dar; die Kampfrichter bewerteten jeweils dieselben Leistungen.
- **kinship_dat.sav.** Rosenberg und Kim (Rosenberg als auch Kim, 1975) haben 15 Bezeichnungen für den Verwandtschaftsgrad untersucht (Tante, Bruder, Cousin, Tochter, Vater, Enkelin, Großvater, Großmutter, Enkel, Mutter, Neffe, Nichte, Schwester, Sohn, Onkel). Die beiden Analytiker baten vier Gruppen von College-Studenten (zwei weibliche und zwei männliche Gruppen), diese Bezeichnungen auf der Grundlage der Ähnlichkeiten zu sortieren. Zwei Gruppen (eine weibliche und eine männliche Gruppe) wurden gebeten, die Bezeichnungen zweimal zu sortieren; die zweite Sortierung sollte dabei nach einem anderen Kriterium erfolgen als die erste. So wurden insgesamt sechs "Quellen" erzielt. Jede Quelle entspricht einer Ähnlichkeitsmatrix mit 15×15 Elementen. Die Anzahl der Zellen ist dabei gleich der Anzahl der Personen in einer Quelle minus der Anzahl der gemeinsamen Platzierungen der Objekte in dieser Quelle.
- **kinship_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine dreidimensionale Lösung für *kinship_dat.sav*.
- **kinship_var.sav.** Diese Datendatei enthält die unabhängigen Variablen *gender* (Geschlecht), *gener*(Generation) und *degree* (Verwandtschaftsgrad), die zur Interpretation der Dimensionen einer Lösung für *kinship_dat.sav* verwendet werden können. Insbesondere können sie verwendet werden, um den Lösungsraum auf eine lineare Kombination dieser Variablen zu beschränken.
- **mailresponse.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines Bekleidungsherstellers geht, der ermitteln möchte, ob die Verwendung von Briefsendungen für das Direktmarketing zu schnelleren Antworten führt als Postwurfsendungen. Die Mitarbeiter in der Bestellannahme erfassen, wie vielen Wochen nach der Postsendung die einzelnen Bestellungen aufgegeben wurden.
- **marketvalues.sav.** Diese Datendatei betrifft Hausverkäufe in einem Neubaugebiet in Algonquin, Illinois, in den Jahren 1999–2000. Diese Verkäufe sind in Grundbucheinträgen dokumentiert.
- **mutualfund.sav.** Diese Datendatei betrifft Aktienmarktdaten für verschiedene Technologieaktien, die in im Index S&P 500 verzeichnet sind. Jeder Fall entspricht einem Unternehmen.
- **nhis2000_subset.sav.** Die "National Health Interview Survey (NHIS)" ist eine große, bevölkerungsbezogene Umfrage in unter der US-amerikanischen Zivilbevölkerung. Es werden persönliche Interviews in einer landesweit repräsentativen Stichprobe von Haushalten durchgeführt. Für die Mitglieder jedes Haushalts werden demografische Informationen und

Beobachtungen zum Gesundheitsverhalten und Gesundheitsstatus eingeholt. Diese Datendatei enthält eine Teilmenge der Informationen aus der Umfrage des Jahres 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Datendatei und Dokumentation öffentlich zugänglich. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Zugriff erfolgte 2003.

- **ozone.sav.** Die Daten enthalten 330 Beobachtungen zu sechs meteorologischen Variablen zur Vorhersage der Ozonkonzentration aus den übrigen Variablen. Bei früheren Untersuchungen ((Breiman als auch Friedman, 1985), (Hastie als auch Tibshirani, 1990)) fanden Wissenschaftler einige Nichtlinearitäten unter diesen Variablen, die die Standardverfahren bei der Regression behindern.
- **pain_medication.sav.** Diese hypothetische Datendatei enthält die Ergebnisse eines klinischen Tests für ein entzündungshemmendes Medikament zur Schmerzbehandlung bei chronischer Arthritis. Von besonderem Interesse ist die Zeitdauer, bis die Wirkung des Medikaments einsetzt und wie es im Vergleich mit bestehenden Medikamenten abschneidet.
- **patient_los.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen zu Patienten, die wegen des Verdachts auf Herzinfarkt in das Krankenhaus eingeliefert wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **patlos_sample.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen für eine Stichprobe von Patienten, denen während der Behandlung eines Herzinfarkts Thrombolytika verabreicht wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **polishing.sav.** Hierbei handelt es sich um die Datendatei "Nambeware Polishing Times" aus der Data and Story Library. Sie bezieht sich auf die Bemühungen eines Herstellers von Metallgeschirr (Nambe Mills, Santa Fe, New Mexico) zur zeitlichen Planung seiner Produktion. Jeder Fall entspricht einem anderen Artikel in der Produktpalette. Für jeden Artikel sind Durchmesser, Polierzeit, Preis und Produkttyp erfasst.
- **poll_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die öffentliche Unterstützung für einen Gesetzentwurf zu ermitteln, bevor er im Parlament eingebracht wird. Die Fälle entsprechen registrierten Wählern. Für jeden Fall sind County, Gemeinde und Wohnviertel des Wählers erfasst.
- **poll_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *poll_cs.sav* aufgeführten Wähler. Die Stichprobe wurde gemäß dem in der Plandatei *poll.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Beachten Sie jedoch Folgendes: Da im Stichprobenplan die PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*poll_jointprob.sav*). Die zusätzlichen Variablen zum demografischen Hintergrund der Wähler und ihrer Meinung zum vorgeschlagenen Gesetzentwurf wurden nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **property_assess.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen Bezirk (County) zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien, die im vergangenen Jahr in dem betreffenden County verkauft wurden. Jeder Fall in der Datendatei enthält die

Gemeinde, in der sich die Immobilie befindet, den Bewerter, der die Immobilie besichtigt hat, die seit dieser Bewertung verstrichene Zeit, den zu diesem Zeitpunkt ermittelten Wert sowie den Verkaufswert der Immobilie.

- **property_assess_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen US-Bundesstaat zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien in dem betreffenden Bundesstaat. Jeder Fall in der Datendatei enthält das County, die Gemeinde und das Wohnviertel, in dem sich die Immobilie befindet, die seit der letzten Bewertung verstrichene Zeit sowie zu diesem Zeitpunkt ermittelten Wert.
- **property_assess_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *property_assess_cs.sav* aufgeführten Immobilien. Die Stichprobe wurde gemäß dem in der Plandatei *property_assess.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Die zusätzliche Variable *Current value* (Aktueller Wert) wurde nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **recidivism.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem frühren Straftäter und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Zeit bis zu seiner zweiten Festnahme, sofern diese innerhalb von zwei Jahren nach der ersten Festnahme erfolgte.
- **recidivism_cs_sample.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter, der im Juni 2003 erstmals aus der Haft entlassen wurde, und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Daten zu seiner zweiten Festnahme, sofern diese bis Ende Juni 2006 erfolgte. Die Straftäter wurden aus per Stichprobenziehung ermittelten Polizeidirektionen ausgewählt (gemäß dem in *recidivism_cs.csplan* angegebenen Stichprobenplan). Da hierbei eine PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Eine hypothetische Datendatei mit Kauftransaktionsdaten wie Kaufdatum, gekauften Artikeln und Geldbetrag für jede Transaktion.
- **salesperformance.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung von zwei neuen Verkaufsschulungen geht. 60 Mitarbeiter, die in drei Gruppen unterteilt sind, erhalten jeweils eine Standardschulung. Zusätzlich erhält Gruppe 2 eine technische Schulung und Gruppe 3 eine Praxisschulung. Die einzelnen Mitarbeiter wurden am Ende der Schulung einem Test unterzogen und die erzielten Punkte wurden erfasst. Jeder Fall in der Datendatei stellt einen Lehrgangsteilnehmer dar und enthält die Gruppe, der der Lehrgangsteilnehmer zugeteilt wurde sowie die von ihm in der Prüfung erreichte Punktzahl.
- **satisf.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Zufriedenheitsumfrage, die von einem Einzelhandelsunternehmen in 4 Filialen durchgeführt wurde. Insgesamt wurden 582 Kunden befragt. Jeder Fall gibt die Antworten eines einzelnen Kunden wieder.

- **screws.sav.** Diese Datendatei enthält Informationen zu den Eigenschaften von Schrauben, Bolzen, Muttern und Reißnägeln ((Hartigan, 1975)).
- **shampoo_ph.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik für Haarpflegeprodukte geht. In regelmäßigen Zeitabständen werden Messwerte von sechs separaten Ausgangschargen erhoben und ihr pH-Wert erfasst. Der Zielbereich ist 4,5–5,5.
- **ships.sav.** Ein an anderer Stelle ((McCullagh et al., 1989)) vorgestelltes und analysiertes Daten-Set bezieht sich auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfallshäufigkeiten können unter Angabe von Schiffstyp, Konstruktionszeitraum und Betriebszeitraum gemäß einer Poisson-Rate modelliert werden. Das Aggregat der Betriebsmonate für jede Zelle der durch die Kreuzklassifizierung der Faktoren gebildeten Tabelle gibt die Werte für die Risikoanfälligkeit an.
- **site.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, neue Standorte für die betriebliche Expansion auszuwählen. Das Unternehmen beauftragte zwei Berater unabhängig voneinander mit der Bewertung der Standorte. Neben einem umfassenden Bericht gaben die Berater auch eine zusammenfassende Wertung für jeden Standort als "good" (gut) "fair" (mittelmäßig) oder "poor" (schlecht) ab.
- **siteratings.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Betatests der neuen Website eines E-Commerce-Unternehmens geht. Jeder Fall entspricht einem Beta-Tester, der die Brauchbarkeit der Website auf einer Skala von 0 bis 20 bewertete.
- **smokers.sav.** Diese Datendatei wurde aus der Umfrage "National Household Survey of Drug Abuse" aus dem Jahr 1998 abstrahiert und stellt eine Wahrscheinlichkeitsstichprobe US-amerikanischer Haushalte dar. Daher sollte der erste Schritt bei der Analyse dieser Datendatei darin bestehen, die Daten entsprechend den Bevölkerungstrends zu gewichten.
- **smoking.sav.** Hierbei handelt es sich um eine von Greenacre ((Greenacre, 1984)) vorgestellte hypothetische Tabelle. Die relevante Tabelle wird durch eine Kreuztabelle der Rauchgewohnheiten und der Berufskategorie gebildet. Die Variable *Berufsgruppe* enthält die Berufskategorien *Senior Manager*, *Junior Manager*, *Angestellter mit Erfahrung*, *Angestellter ohne Erfahrung* und *Sekretariat* sowie die Kategorie *National Average*, die als Ergänzung der Analyse dienen kann. Die Variable *Rauchen* enthält die Rauchgewohnheiten *Nichtraucher*, *Leicht*, *Mittel* und *Stark* sowie die Kategorien *No Alcohol* und *Alcohol*, die als Ergänzung der Analyse dienen können.
- **storebrand.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Verkaufsleiterin in einem Lebensmittelmarkt geht, die die Verkaufszahlen des Waschmittels der Eigenmarke gegenüber den anderen Marken steigern möchte. Sie erarbeitet eine Werbeaktion im Geschäft und spricht an der Kasse mit Kunden. Jeder Fall entspricht einem Kunden.
- **stores.sav.** Diese Datendatei enthält hypothetische monatliche Marktanteilsdaten für zwei konkurrierende Lebensmittelgeschäfte. Jeder Fall entspricht den Marktanteilsdaten für einen bestimmten Monat.
- **stroke_clean.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozeduren in der Option "Data Preparation" bereinigt wurde.

- **stroke_invalid.sav.** Diese hypothetische Datendatei enthält den ursprünglichen Zustand einer medizinischen Datenbank, der mehrere Dateneingabefehler aufweist.
- **stroke_survival.** In dieser hypothetischen Datendatei geht es um die Überlebenszeiten von Patienten, die nach einem Rehabilitationsprogramm wegen eines ischämischen Schlaganfalls mit einer Reihe von Problemen zu kämpfen haben. Nach dem Schlaganfall werden das Auftreten von Herzinfarkt, ischämischem Schlaganfall und hämorrhagischem Schlaganfall sowie der Zeitpunkt des Ereignisses aufgezeichnet. Die Stichprobe ist auf der linken Seite abgeschnitten, da sie nur Patienten enthält, die bis zum Ende des Rehabilitationprogramms, das nach dem Schlaganfall durchgeführt wurde, überlebten.
- **stroke_valid.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozedur "Daten validieren" überprüft wurde. Sie enthält immer noch potenziell anomale Fälle.
- **survey_sample.sav.** Diese hypothetische Datendatei enthält Umfragedaten einschließlich demografischer Daten und verschiedener Meinungskennzahlen.
- **tastetest.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung der Auswirkungen der Mulchfarbe auf den Geschmack von Pflanzenprodukten geht. Der Geschmack von Erdbeeren, die in rotem, blauem und schwarzem Rindenmulch gezogen wurden, wurde von Testpersonen auf einer ordinalen Skala (weit unter bis weit über dem Durchschnitt) bewertet. Jeder Fall entspricht einem Geschmackstester.
- **telco.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Telekommunikationsunternehmens geht, die Kundenabwanderung zu verringern. Jeder Fall entspricht einem Kunden und enthält verschiedene Informationen zum demografischen Hintergrund und zur Servicenutzung.
- **telco_extra.sav.** Diese Datendatei ähnelt der Datei *telco.sav*, allerdings wurden die Variablen "tenure" und die Log-transformierten Variablen zu den Kundenausgaben entfernt und durch standardisierte Log-transformierte Variablen ersetzt.
- **telco_missing.sav.** Diese Datendatei ist eine Untermenge der Datendatei *telco.sav*, allerdings wurde ein Teil der demografischen Datenwerte durch fehlende Werte ersetzt.
- **testmarket.sav.** Diese hypothetische Datendatei bezieht sich auf die Pläne einer Fast-Food-Kette, einen neuen Artikel in ihr Menü aufzunehmen. Es gibt drei mögliche Kampagnen zur Verkaufsförderung für das neue Produkt. Daher wird der neue Artikel in Filialen in mehreren zufällig ausgewählten Märkten eingeführt. An jedem Standort wird eine andere Form der Verkaufsförderung verwendet und die wöchentlichen Verkaufszahlen für das neue Produkt werden für die ersten vier Wochen aufgezeichnet. Jeder Fall entspricht einer Standort-Woche.
- **testmarket_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *testmarket.sav*, wobei die wöchentlichen Verkaufszahlen zusammengefasst sind, sodass jeder Fall einem Standort entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und die verzeichneten Verkaufszahlen sind nun die Summe der Verkaufszahlen während der vier Wochen der Studie.
- **tree_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_credit.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält.

- **tree_missing_data.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält und eine große Anzahl fehlender Werte aufweist.
- **tree_score_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_textdata.sav.** Eine einfache Datendatei mit nur zwei Variablen, die vor allem den Standardzustand von Variablen vor der Zuweisung von Messniveau und Wertelabels zeigen soll.
- **tv-survey.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Studie, die von einem Fernsehstudio durchgeführt wurde, das überlegt, ob die Laufzeit eines erfolgreichen Programms verlängert werden soll. 906 Personen wurden gefragt, ob sie das Programm unter verschiedenen Bedingungen ansehen würden. Jede Zeile entspricht einem Befragten; jede Spalte entspricht einer Bedingung.
- **ulcer_recurrence.sav.** Diese Datei enthält Teilinformationen aus einer Studie zum Vergleich der Wirksamkeit zweier Therapien zur Vermeidung des Wiederauftretens von Geschwüren. Es stellt ein gutes Beispiel für intervallzensierte Daten dar und wurde an anderer Stelle ((Collett, 2003)) vorgestellt und analysiert.
- **ulcer_recurrence_recoded.sav.** In dieser Datei sind die Daten aus *ulcer_recurrence.sav* so umstrukturiert, dass das Modell der Ereigniswahrscheinlichkeit für jedes Intervall der Studie berechnet werden kann und nicht nur die Ereigniswahrscheinlichkeit am Ende der Studie. Sie wurde an anderer Stelle ((Collett et al., 2003)) vorgestellt und analysiert.
- **verd1985.sav.** Diese Datendatei enthält eine Umfrage ((Verdegaal, 1985)). Die Antworten von 15 Subjekten auf 8 Variablen wurden aufgezeichnet. Die relevanten Variablen sind in drei Sets unterteilt. Set 1 umfasst *alter* und *heirat*, Set 2 besteht aus *pet* und *news* und in Set 3 finden sich *music* und *live*. Die Variable *pet* wird mehrfach nominal skaliert und die Variable *Alter* ordinal. Alle anderen Variablen werden einzeln nominal skaliert.
- **virus.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Internet-Dienstanbieters geht, der die Auswirkungen eines Virus auf seine Netzwerke ermitteln möchte. Dabei wurde vom Moment der Virusentdeckung bis zu dem Zeitpunkt, zu dem die Virusinfektion unter Kontrolle war, der (ungefähre) prozentuale Anteil infizierter E-Mail in den Netzwerken erfasst.
- **waittimes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Wartezeiten für Kunden bei drei verschiedenen Filialen einer Bank. Jeder Fall entspricht einem Kunden und zeichnet die Wartezeit und die Filiale.
- **webusability.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests zur Benutzerfreundlichkeit eines neuen Internetgeschäfts geht. Jeder Fall entspricht einer von fünf Testpersonen, die die Benutzerfreundlichkeit bewerten und gibt für sechs separate Aufgaben an, ob die Testperson sie erfolgreich ausführen könnte.
- **wheeze_steubenville.sav.** Hierbei handelt es sich um eine Teilmenge der Daten aus einer Langzeitstudie zu den gesundheitlichen Auswirkungen der Luftverschmutzung auf Kinder ((Ware, Dockery, Spiro III, Speizer, als auch Ferris Jr., 1984)). Die Daten enthalten wiederholte binäre Messungen des Keuchens von Kindern aus Steubenville, Ohio, im Alter

von 7, 8, 9 und 10 Jahren sowie eine unveränderlichen Angabe, ob die Mutter im ersten Jahr der Studie rauchte oder nicht.

- **workprog.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einem Arbeitsprogramm der Regierung, das versucht, benachteiligten Personen bessere Arbeitsplätze zu verschaffen. Eine Stichprobe potenzieller Programmteilnehmer wurde beobachtet. Von diesen Personen wurden nach dem Zufallsprinzip einige für die Teilnahme an dem Programm ausgewählt. Jeder Fall entspricht einem Programmteilnehmer.

Bibliografie

Bell, U. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Blake, R. L., als auch R. O. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., als auch O. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580–598.

Collett, D. 2003. *Modelling survival data in medical research*, 2 (Hg.). Boca Raton: Chapman & Hall/CRC.

Green, K. U., als auch G. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, K. U., als auch J. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Greenacre, A. O. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, 469–506.

Hartigan, O. D. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, B., als auch L. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Kennedy, L., R. Riquier, als auch K. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, 56–70.

McCullagh, P., als auch O. D. Nelder. 1989. *Verallgemeinerte lineare Modelle*, 2nd (Hg.). London: Chapman & Hall.

Menec , V., N. Roos, D. Nowicki, L. MacWilliam, G. Finlayson , als auch C. Schwarz. 1999. *Seasonal Patterns of Winnipeg Hospital Use*. : Manitoba Centre for Health Policy.

Price, L. H., als auch P. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, 579–586.

Rickman, L., X. Mitchell, O. Dingman, als auch O. U. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, 54–58.

Rosenberg, F., als auch A. K. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, 489–502.

Van der Ham, B., O. O. Meulman, P. R. Van Strien, als auch H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, 363–368.

Verdegaal, L. 1985. *Meer sets analyse voor kwalitatieve gegevens (in niederländischer Sprache)*. Leiden: Department of Data Theory, Universität Leiden.

Ware, O. H., P. F. Dockery, D. Spiro III, U. U. Speizer, als auch K. A. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, 366–374.

Index

Analysegewichtung
in der automatisierten Datenaufbereitung, 25
Anomalie-Indizes
in “Ungewöhnliche Fälle identifizieren”, 47–48, 114
Automatische Datenaufbereitung, 17
Automatisierte Datenaufbereitung, 82
Aktionsdetails, 40
Aktionsübersicht, 35
Ansichten zurücksetzen, 32
automatisch, 95
Datenqualität verbessern, 24
Datum und Uhrzeit aufbereiten, 21
Feldanalyse, 33
Felddetails, 38, 90
Felder, 20
Felder ausschließen, 22
Felder neu skalieren, 25
Felder transformieren, 26
Feldertabelle, 37
Feldverarbeitungsübersicht, 32
Funktionsauswahl, 27
Funktionserstellung, 27
interaktiv, 82
Messniveau anpassen, 23
Modellansicht, 31
Namensfelder, 28
Stetiges Ziel normalisieren, 25
Transformationen anwenden, 29
Verknüpfungen zwischen Ansichten, 32
Vorhersagekraft, 36
Werte zurücktransformieren, 43
Ziele, 17

Beispieldateien
Lage, 139
Binning-Regeln
in “Optimales Klassieren”, 55
Box-Cox-Transformation
in der automatisierten Datenaufbereitung, 25

Daten validieren, 8, 60
Ausgabe, 14
Fallbericht, 72, 80
Gleiche Fallbezeichner, 63
grundlegende Prüfungen, 10
Regelbeschreibung, 71
Regeln für eine Variable, 12
Regeln für mehrere Variablen, 13, 79
Unvollständige Fallbezeichner, 63
Variablen speichern, 15
Variablenauswertung, 71
verwandte Prozeduren, 81
Warnungen, 62
Datenvalidierung
in “Daten validieren”, 8
Dauer berechnen
Automatisierte Datenaufbereitung, 21
Dauerberechnung
Automatisierte Datenaufbereitung, 21
Deskriptive Statistiken
in “Optimales Klassieren”, 129

Endpunkte für Klassen
in “Optimales Klassieren”, 54

Fallbericht
in “Daten validieren”, 72, 80
Fehlende Werte
in “Ungewöhnliche Fälle identifizieren”, 49
Felddetails
Automatisierte Datenaufbereitung, 90
Funktionsauswahl
in der automatisierten Datenaufbereitung, 27
Funktionserstellung
in der automatisierten Datenaufbereitung, 27

Gleiche Fallbezeichner
in “Daten validieren”, 15, 63
Gründe
in “Ungewöhnliche Fälle identifizieren”, 47–48, 116, 120
Gruppen
in “Ungewöhnliche Fälle identifizieren”, 47–48, 113, 115

Interaktive Datenaufbereitung, 17

Klassierte Variablen
in “Optimales Klassieren”, 135
Klassierungs-Zusammenfassungen
in “Optimales Klassieren”, 131

Leere Fälle
in “Daten validieren”, 15

MDLP
in “Optimales Klassieren”, 52
Modellansicht
in der automatisierten Datenaufbereitung, 31

Modellentropie
in “Optimales Klassieren”, 130

Normwerte der Gruppen
in “Ungewöhnliche Fälle identifizieren”, 117–118

Optimales Klassieren, 52, 125
Ausgabe, 54
Deskriptive Statistiken, 129
Fehlende Werte, 56
Klassierte Variablen, 135
Klassierungs-Zusammenfassungen, 131
Modell, 125
Modellentropie, 130
Optionen, 57
Speichern, 55
Syntax-Klassierungsregeln, 135

Pre-Binning
in “Optimales Klassieren”, 57

Regelbeschreibung
in “Daten validieren”, 71

Stetiges Ziel normalisieren, 25

Überwachtes Binning
im Vergleich mit unüberwachtem Binning, 52
in “Optimales Klassieren”, 52
Ungewöhnliche Fälle identifizieren, 45, 108
Ausgabe, 47
Auswertung der Gründe, 120
Auswertung des Anomalie-Index, 120
Fehlende Werte, 49
Liste der Gründe anomaler Fälle, 116
Liste der Gruppen-IDs anomaler Fälle, 115
Liste der Indizes anomaler Fälle, 114
Modell, 108
Modelldatei exportieren, 48
Normwerte der kategorialen Variablen, 118
Normwerte der metrischen Variablen, 117
Optionen, 50
Variablen speichern, 48
verwandte Prozeduren, 124
Zusammenfassung der Fallverarbeitung, 113
Unüberwachtes Binning
im Vergleich mit überwachtem Binning, 52
Unvollständige Fallbezeichner
in “Daten validieren”, 15, 63

Validierungsregeln, 2
Validierungsregeln definieren, 3
Regeln für eine Variable, 4
Regeln für mehrere Variablen, 6
Validierungsregeln für eine Variable definieren, 73
in “Daten validieren”, 12
in “Validierungsregeln definieren”, 4
Validierungsregeln für mehrere Variablen definieren, 73
in “Daten validieren”, 13, 79
in “Validierungsregeln definieren”, 6
Validierungsregelverletzungen
in “Daten validieren”, 15
Variablenauswertung
in “Daten validieren”, 71
Verletzungen von Validierungsregeln
in “Daten validieren”, 15

Warnungen
in “Daten validieren”, 62

Zusammenfassung der Fallverarbeitung
in “Ungewöhnliche Fälle identifizieren”, 113
Zyklische Zeitelemente
Automatisierte Datenaufbereitung, 21